JMIRx Med

Overlay journal for preprints with post-review manuscript marketplace
Volume 6 (2025) ISSN 2563-6316 Editor in Chief: Edward Meinert, MA (Oxon), MSc, MBA, MPA, PhD, CEng, FBCS, EUR ING

Contents

Authors’ Response To Peer Reviewss

Original Papers

Effects of Interventions for the Prevention and Management of Maternal Anemia in the Advent of the COVID-19 Pandemic: Systematic Review and Meta-Analysis

John Kyalo Muthuka[1,2], DIP PHARM, HND, BSc, PGD, MPH, PhD; Dianna Kageni Mbari-Fondo[3], PM, BEd, MSc, MPH; Francis Muchiri Wambura[4], HND, BSc, Dip-Med Lab, MPH; Kelly Oluoch[4], BPharm, MPharm, MBA, PhD; Japheth Mativo Nzioki[5], BSc (EVH), BSc (ENSc), CPH, MPH, PhD; Everlyn Musangi Nyamai[4], BScN, MPH, PhD; Rosemary Nabaweesi[6], MPH, MBChB, DrPH

[1]Department of Community Health and Health Promotion, Faculty of Public Health, Kenya Medical Training College, Mbagathi Way, Kenyatta National Hospital, Nairobi, Kenya

[2]Epidemiology/Public Health Section, KEMRI Graduate School of Health Sciences, Kenya Medical Research Institute, Nairobi, Kenya

[3]Alberta Health Services, Edmonton, AB, Canada

[4]Kenya Medical Training College, Nairobi, Kenya

[5]School of Nursing, Andrews University, Berrien Springs, MI, United States

[6]School of Global Health, Meharry Medical College, Nashville, TN, United States

Corresponding Author:
John Kyalo Muthuka, DIP PHARM, HND, BSc, PGD, MPH, PhD
Department of Community Health and Health Promotion, Faculty of Public Health, Kenya Medical Training College, Mbagathi Way, Kenyatta National Hospital, Nairobi, Kenya

Related Articles:

Companion article: https://www.medrxiv.org/content/10.1101/2024.02.18.24302492v1

Companion article: https://med.jmirx.org/2025/1/e81700

Companion article: https://med.jmirx.org/2025/1/e81699

Companion article: https://med.jmirx.org/2025/1/e82836

Companion article: https://med.jmirx.org/2025/1/e81711

Abstract

Background: The COVID-19 pandemic presented many unknowns for pregnant women, with anemia potentially worsening pregnancy outcomes due to multiple factors.

Objective: This review aimed to determine the pooled effect of maternal anemia interventions and associated factors during the pandemic.

Methods: Eligible studies were observational and included reproductive-age women receiving anemia-related interventions during the COVID-19 pandemic. Exclusion criteria comprised non-English publications, reviews, editorials, case reports, studies with insufficient data, sample sizes below 50, and those lacking DOIs. A systematic search of PubMed, Scopus, Embase, Web of Science, and Google Scholar identified articles published between December 2019 and August 2022. Risk of bias was evaluated using the Cochrane Risk of Bias 2 tool for randomized trials and the National Institutes of Health's assessment tool for observational studies. Pooled rate ratios (RRs) with 95% CIs were calculated in Review Manager 5.4.1. Synthesis included subgroup analysis, meta-regression, and publication bias checks to assess intervention effectiveness.

Results: This meta-analysis included 11 studies with 6129 pregnant women. Of these, 3591 (59%) were in the intervention group and 2538 (41%) were in the comparator group. Effects were recorded for 1921 (53.4%) women in the intervention group and 1350 (53.1%) in the comparator group. The cumulative impact ranged from 23% to 81%, averaging 56%. The initial analysis showed no significant effect on anemia prevention (RR 0.79, 95% CI 0.61 - 1.02; P=.07), with high heterogeneity (I^2=97%). Sensitivity analysis excluding 4 outlier studies improved the effect size to a significant level at 39% (RR 0.61, 95% CI 0.43 - 0.87;

P=.006). Subgroup analysis revealed substantial heterogeneity (I^2=87.2%). Intravenous sucrose had a poor impact (RR 1.31, 95% CI 1.17 - 1.47; *P*<.001), while medicinal or herbal interventions showed benefit (RR 0.81, 95% CI 0.73 - 0.90; *P*=.006). Educational interventions yielded a 28% effect (RR 0.72), medicinal administration 19% (RR 0.81), iron supplementation 17% (RR 0.83), and intravenous ferric carboxylmaltose 15% (RR 0.85; *P*<.02). Additional sensitivity analysis confirmed a pooled positive effect of 17% (RR 0.83, 95% CI 0.79 - 0.88; *P*<.001), with minimal heterogeneity (I^2=0%). Regionally, effectiveness was highest in Africa (RR 0.84, 95% CI 0.79 - 0.89; *P*<.001). Multicenter studies and those with 2020 data were predictive of better outcomes (RR 0.84 and RR 0.50, respectively). Despite initial heterogeneity and publication bias, interventions showed utility in mitigating maternal anemia in targeted subgroups and regions.

Conclusions: Maternal anemia interventions during the COVID-19 pandemic showed modest, context-specific effectiveness, with declining impact from 2020 to 2022. Although high heterogeneity and study inconsistencies limited generalizability, significant benefits were observed particularly in African and multicenter studies. The pandemic exposed gaps in maternal health systems, emphasizing the need for tailored interventions, stronger data infrastructure, and resilient care strategies in future global crises.

Trial Registration: PROSPERO CRD42023410657; https://www.crd.york.ac.uk/PROSPERO/view/CRD42023410657

(JMIRx Med 2025;6:e57626) doi:10.2196/57626

KEYWORDS

maternal anemia; anemia in pregnancy; COVID-19; pregnancy complications; meta-analysis; maternal and child health; anemia prevention; reproductive health

Introduction

Anemia is a condition where the number of red blood cells or the hemoglobin concentration within them is lower than normal. Maternal anemia refers to pregnant women having hemoglobin levels less than 12 g/dL [1-3]. Studies have found a correlation between the prevalence of anemia in women and the gross domestic product per capita. Projections suggest a 10% decline in global gross domestic product due to COVID-19, with findings indicating that the availability of nutritious foods, in particular, has been affected by COVID-19 measures [4].

Globally, the COVID-19 pandemic has had devastating effects on health care delivery systems for people of all ages, but pregnant women face particular challenges [5,6]. Reports show that the pandemic is making it increasingly challenging to provide adequate maternity care worldwide [5,7]. Even the movement of people seeking to access health care services has been restricted in many countries to prevent the spread of the virus. The pandemic has led to a complete stoppage of the import and export of many essential commodities among various countries, leading to a shortage of necessary items and affecting health care services badly, especially sexual and reproductive health care [8,9]. The population was advised not to go to hospitals unless strictly necessary; this advice seems to apply to all, including healthy pregnant women and even those with complications [5,10].

Before COVID-19, anemia prevention interventions focused on iron and folic acid supplementation, dietary modifications, and public health campaigns [11-13]. During the COVID-19 pandemic, these interventions adapted to include telemedicine, remote consultations, and increased community health worker involvement to address health care disruptions [14-17]. These measures aimed to ensure continued support for pregnant women [18,19]. Interventions to prevent anemia in pregnant women included iron and folic acid supplementation, dietary modifications, education and awareness programs, telemedicine and remote consultations, and community-based interventions [20,21]. The World Health Assembly set 6 targets to be accomplished by the year 2025. Among the targets is a 50% reduction of anemia in women of reproductive age through several strategies such as food fortification with iron, folic acid, and other micronutrients; the distribution of iron-containing supplements; and the control of infections and malaria [22].

There were many unknowns for pregnant women during the COVID-19 pandemic. Some issues may have gone unnoticed; however, conditions such as anemia could lead to worse pregnancy outcomes. Standard intervention efforts may have been compromised due to the COVID-19 pandemic, as was reported during prior pandemics, affecting the effect of health interventions in vulnerable populations [23].

The COVID-19 pandemic has had significant direct and indirect effects on pregnant women, newborns, young children, and adolescents. Directly, pregnant women infected with COVID-19 faced increased risks of preterm birth and stillbirth. However, the transmission of the virus from pregnant women to their newborns was found to be very low [10,24-26]. Indirectly, the pandemic led to reduced prenatal care visits, strained health care infrastructure, and increased maternal mental health issues such as anxiety and depression. Additionally, social and economic disruptions caused by the pandemic exacerbated domestic violence and financial instability, disproportionately affecting women and children [10].

These combined effects highlight the need for targeted interventions to support maternal and child health during and after the pandemic. The effects on pregnant women, newborn babies, young children, and adolescents are enormous and possibly translate to interventions meant to mitigate anemic conditions in pregnancy [6,7,9,10,27-29]. The objective of the review was to assess the cumulative impact of interventions for maternal anemia and related factors during the COVID-19 pandemic.

Methods

Design

All guidelines listed in the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) statement were followed in performing this meta-analysis [30]. For this systematic review and meta-analysis, data were pooled from observational studies, including cohort, case-control, cross-sectional, and similar viable case studies. The study was registered on PROSPERO (International Prospective Register of Systematic Reviews; CRD42023410657).

Search Strategy

We performed a simple search in the Google Scholar, PubMed, Scopus, Web of Science, and Embase databases to identify observational studies suitable for inclusion with the following search terms: “maternal anemia” OR “anemic condition” OR “poor hemoglobin levels” OR “pregnancy anemia” OR “anemia in pregnant women” OR “gestation anemia” AND “treatment” OR “intervention” OR “management” AND “effect” OR “effectiveness” AND “impact” OR “outcome.” Studies were restricted to those published in English from December 2019 to August 2022.

Inclusion and Exclusion Criteria

The inclusion criteria for this study were as follows:

1. Studies that examined women of reproductive age who were part of any anemia prevention program or intervention, whether they were anemic or nonanemic according to World Health Organization criteria.
2. Observational, cross-sectional, prospective, or retrospective studies.
3. Studies that compared intervention approaches with control or comparator approaches.
4. Studies evaluating the effects of different interventions on pregnant women during the advent of the COVID-19 pandemic.

The exclusion criteria for this study were as follows:

1. Unrelated, duplicate, and missing information answering our research question.
2. Non–English-language studies.
3. Case reports/series.
4. Reviews.
5. Editorials.
6. Studies lacking a full text (unavailable or not yet published).
7. Articles without a DOI.
8. Studies with small sample sizes (<50 patients), due to low statistical power.

Data Extraction

Both adjusted and nonadjusted data for pregnant women receiving interventions versus those in the comparator group were extracted to identify the most relevant confounding factors for subsequent pooling analysis. Two reviewers (JKM and DMF) scanned study titles and abstracts obtained from the initial database search and included relevant articles in a secondary pool. Next, two independent reviewers (FMW and KO) evaluated the full texts of these articles to determine if they met the study inclusion criteria. Any disputes were resolved through discussion and negotiation with a fourth reviewer (EMN). Only studies agreed upon by all reviewers were included in the final analysis.

The following data were obtained from all studies: title, first author, data collection year, region, sample size, study design, study setting (single or multicenter), intervention type, and the effect associated with each intervention approach. The analysis aimed to determine whether the intervention group was more likely to experience a better effect on maternal anemia mitigation, treatment, or management using end-result indicators such as hemoglobin levels. Further sensitivity and subgroup analyses were also conducted.

Risk of Bias (Quality) Assessment

To assess the quality of randomized controlled trials (RCTs), the Cochrane Risk of Bias tool and Risk of Bias 2 tool [31] were used, evaluating domains such as randomization, deviations from intended interventions, missing outcome data, outcome measurement, and selection of reported results. For observational and cross-sectional studies, the National Institutes of Health tool was used [32]. Two to three reviewers independently assessed study quality, rating each of the 14 items as yes, no, or not applicable. Overall scores were calculated to classify studies as poor, fair, or good. To reduce bias, data checks were performed by reviewers who did not initially extract the data, though some overlap occurred in rare cases.

Statistical Analyses

Review Manager 5.4.1 was used to calculate rate ratios (RRs) with 95% CIs, depicted using forest plots. Quantitative variables were summarized as total numbers and percentages. RRs that did not favor the intervention arm were noted. The effects on anemia prevention, control, management, and treatment were compared between intervention and control arms.

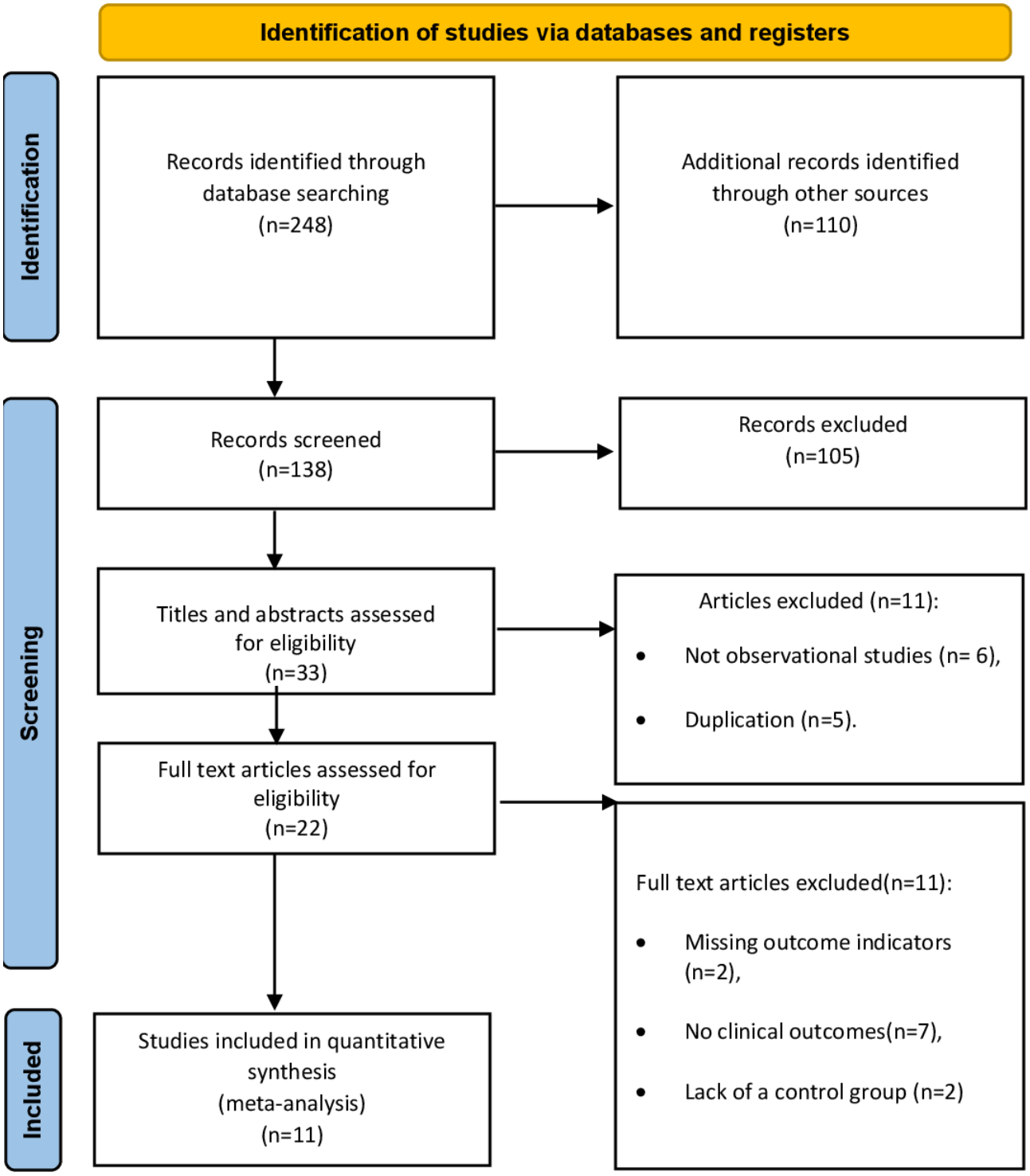
Heterogeneity was evaluated using the Cochran Q and Higgins tests, applying fixed-effects or random-effects models based on heterogeneity levels. Sensitivity adjustments were made to identify sources of heterogeneity by excluding studies one at a time. Subgroup analysis, cumulative analyses, and meta-regression were performed to test result consistency and the impact of confounders on anemia control. Publication bias was assessed using the Cochrane Risk of Bias tool.

Results

Included Articles and Quality Assessment

The initial search of international databases using the specified keywords yielded 248 articles. After excluding 110 duplicates, 138 articles remained. Upon evaluating the titles and abstracts for appropriateness, 33 articles met the inclusion criteria. Additionally, 22 articles were excluded after full-text review for not meeting the inclusion criteria. Ultimately, 11 articles met the inclusion criteria [33-43] (Figure 1).

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flowchart of the study selection procedures.



Features of the Included Studies

The 11 included studies provided data for 6129 pregnant women in the advent of the COVID-19 pandemic [33-43]. Among the 6129 pregnant women included in the meta-analysis, 3591 (59%) were in the maternal anemia intervention group and 2538 (41%) were in the comparator group. The effects of the intervention were reported for 1921 participants (53.4%) in the intervention group and 1350 participants (53.1%) in the comparator group. The cumulative effect on maternal anemia for both groups ranged from 23% to 81%, with an average of 56%. The main outcome of this meta-analysis was the pooled effect of interventions on maternal anemia, assessed by increased hemoglobin levels and other parameters. The study designs included 4 RCTs (2 multicenter, 2 single-center), 3 cross-sectional studies (all multicenter), 2 prospective studies (1 multicenter, 1 single-center), 1 retrospective case-control study (single-center), and 1 quasi-experimental study (single-center). A summary of the studies is provided in Table S1 in Multimedia Appendix 1.

We evaluated the quality of observational studies using a modified Newcastle-Ottawa Scale, which includes 8 items across 3 subscales. Studies scoring ≥7 were considered high quality,

though no universal standard exists. Out of 11 studies, the average score was 6.7, indicating moderate quality (score range: 5 - 8; see Table S2 in Multimedia Appendix 1).

The Pooled Effect of Interventions on the Prevention and Management of Maternal Anemia

The meta-analysis revealed a nonsignificant effect of the interventions on the prevention and management of maternal anemia as indicated by stabilized hemoglobin levels and other parameters (random-effects model RR 0.79, 95% CI 0.61 - 1.02; P=.07; χ^2_{10}=286.98, P<.001; I^2=97%). Based on the confidence interval, this indicated little knowledge about the effect and this imprecision affected the certainty in the evidence; thus, further information was needed before a more certain conclusion could be made (Figure 2). A funnel plot demonstrated an asymmetrical shape, depicting the presence of publication bias (Figure 3).

Figure 2. A forest plot of a meta-analysis of the effect of maternal anemia interventions [33-43]. M-H: Mantel-Haenszel.

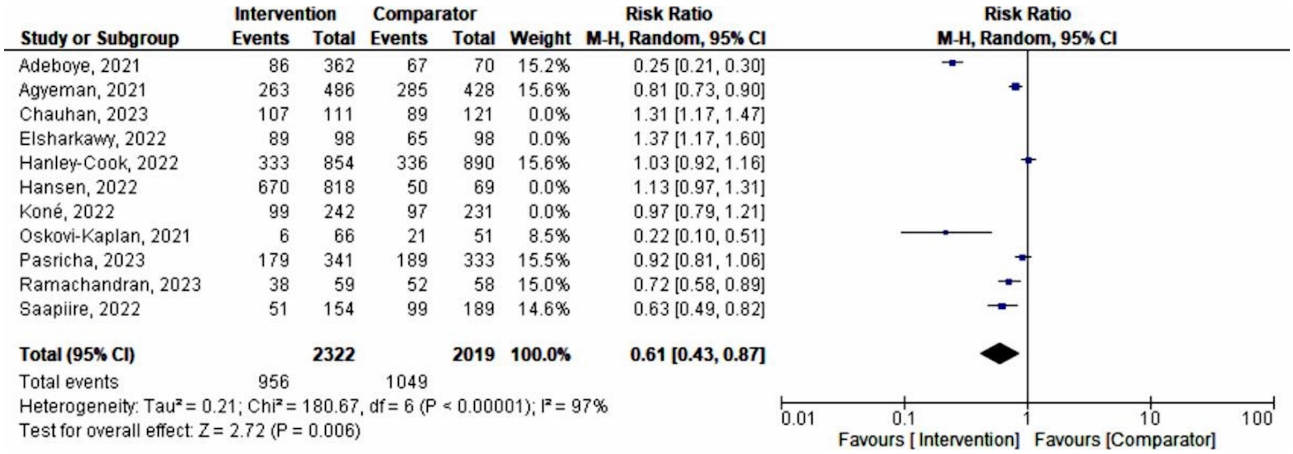
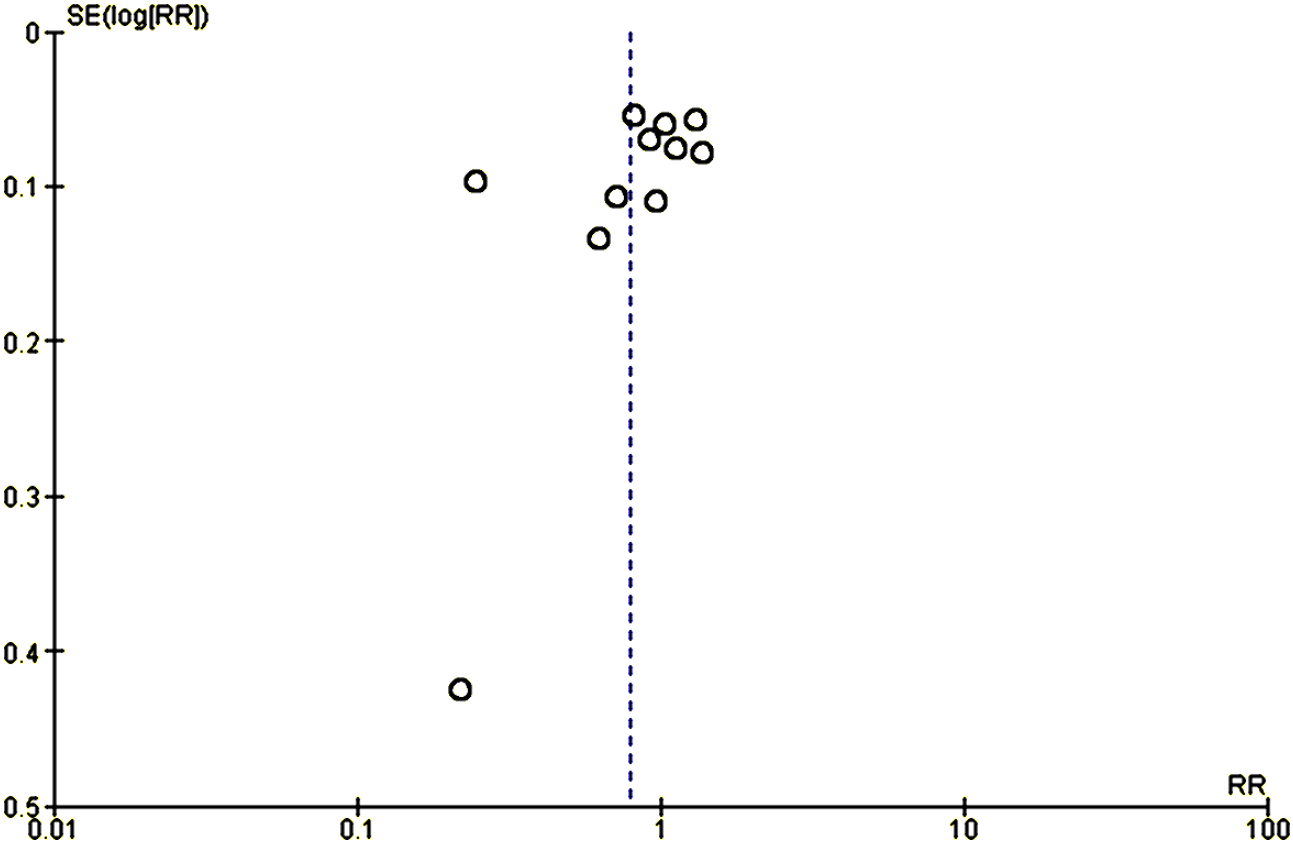


Figure 3. Funnel plot evaluating publication bias. RR: relative risk; SE: standard error.



A sensitivity analysis was performed to explore the impact of excluding or including studies in the meta-analysis based on sample size, methodological quality, and variance. After removing 4 studies (with 1788 pregnant women) [32-34,36] with wider 95% CIs, a total of 4341 pregnant women remained for analysis in the remaining studies [29-31,35,37-39], showing a shift in a random effects model (RR 0.61, 95% CI 0.43 - 0.87; P=.006; χ^2_6=286.98, P<.001; I^2=97%), revealing that the interventions had a 39% utility in preventing and managing maternal anemia during the advent of the COVID-19 pandemic (Figure 4). The funnel plot evaluating publication bias revealed

considerable heterogeneity between all pooled studies for the updated analysis (I^2=97%; P<.001; Figure 5).

Figure 4. A forest plot of a meta-analysis on the effect of maternal anemia interventions after sensitivity analysis [33-43]. M-H: Mantel-Haenszel.

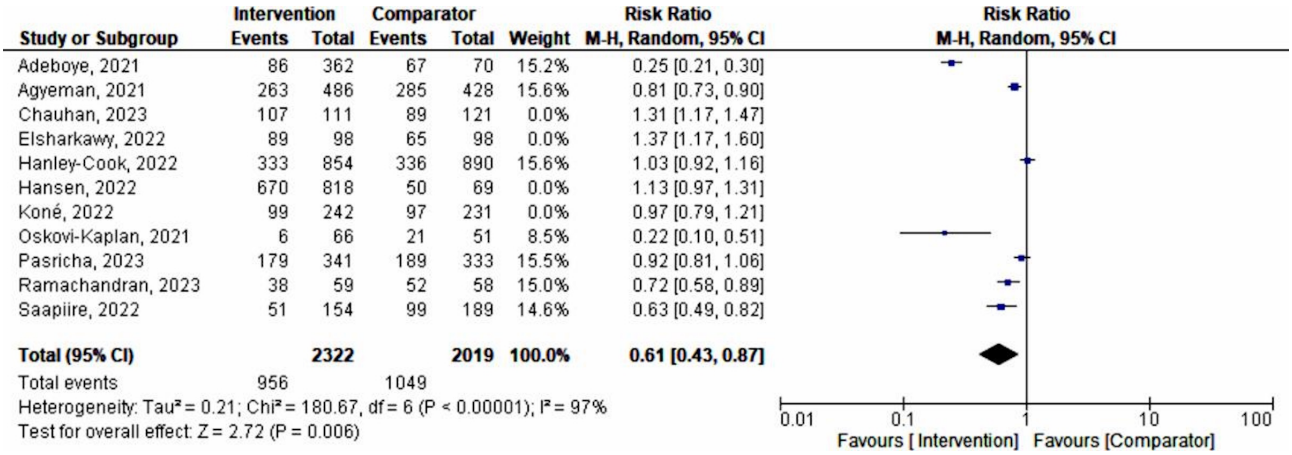
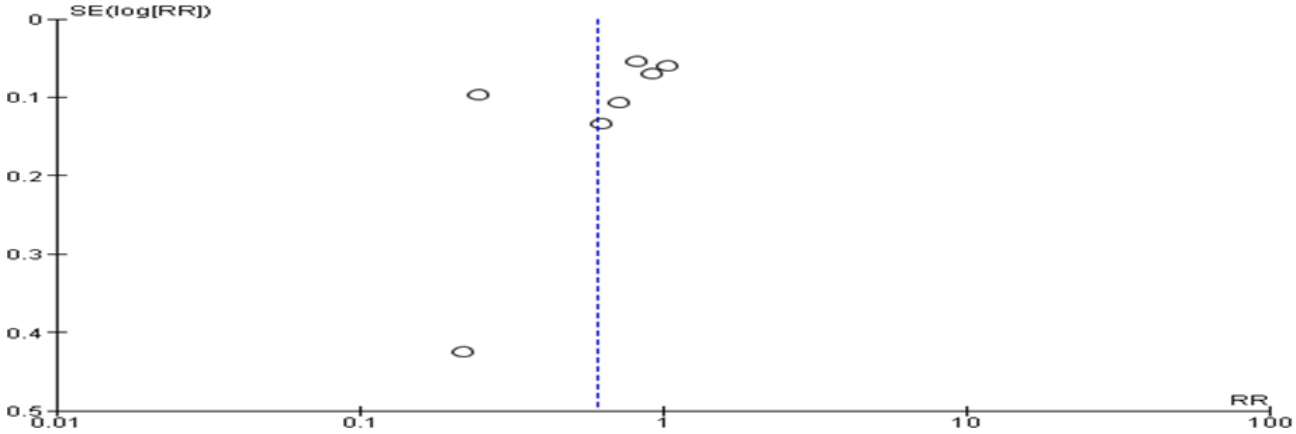


Figure 5. Funnel plot evaluating publication bias after sensitivity analysis. RR: rate ratio; SE: standard error.



Subgroup Analysis and Investigation of Heterogeneity

Heterogeneity in the pooled effect estimates was considerably high for all 11 studies, with 1788 of 6129 (29%) evaluated subjects contributing to this variability. Therefore, it was necessary to perform subgroup analyses to identify possible variables or characteristics moderating the results.

Subgroup analysis with a random-effects model was conducted according to the type or form of the intervention used, including dietary iron supplementation (n=2176), education or dietary information (n=786), intravenous (IV) ferric carboxylmaltose (n=791), medicinal or herbal administration (n=914), IV sucrose (n=232), and other forms (n=1230). This analysis still showed considerable heterogeneity (χ^2 =38.92, P<.001; I^2=87.2%).

The tests for the overall effect of dietary iron supplementation (z=0.92, P=.36), education or dietary information (z=.05, P=.96), ferric carboxylmaltose (z=1.01, P=.31), and other interventions (z=0.51; P=.61) all indicated no significant difference, with substantial heterogeneity (I^2>90%).

Intravenous sucrose (RR 1.31, 95% CI 1.17-1.47; z=4.70; P<.001) demonstrated poor prevention and management of maternal anemia, favoring the comparator by 31%. Meanwhile, medicinal or herbal administration had a 19% effect on the prevention and management of maternal anemia (random-effects model RR 0.81, 95% CI 0.73-0.90; P=.006; Figure 6). Publication bias was further demonstrated by a funnel plot (Figure 7).

Figure 6. Subgroup analysis according to the type or form of intervention, showing similarly high heterogeneity as compared with the full meta-analysis [33-43]. IV: intravenous; M-H: Mantel-Haenszel.

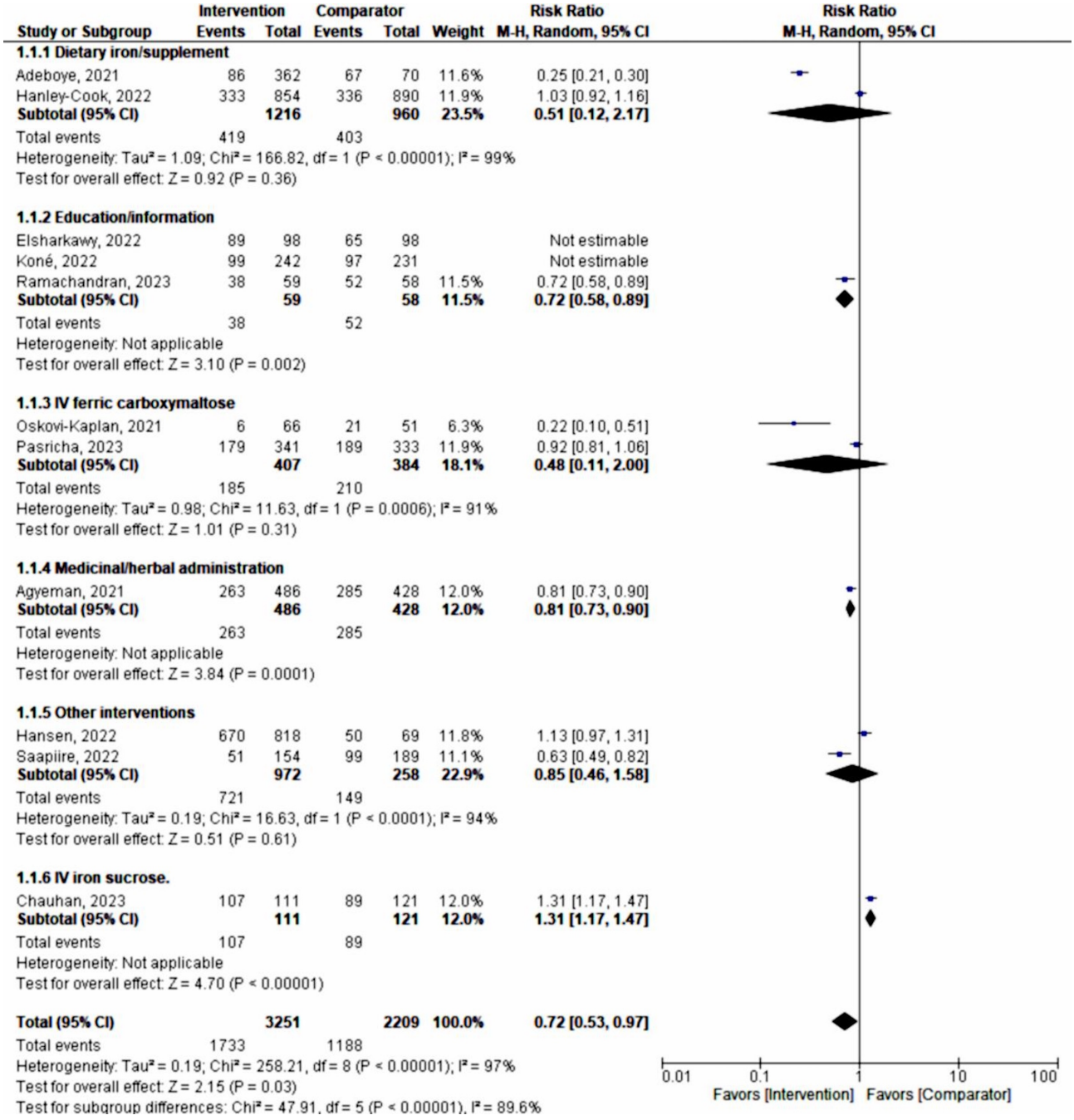
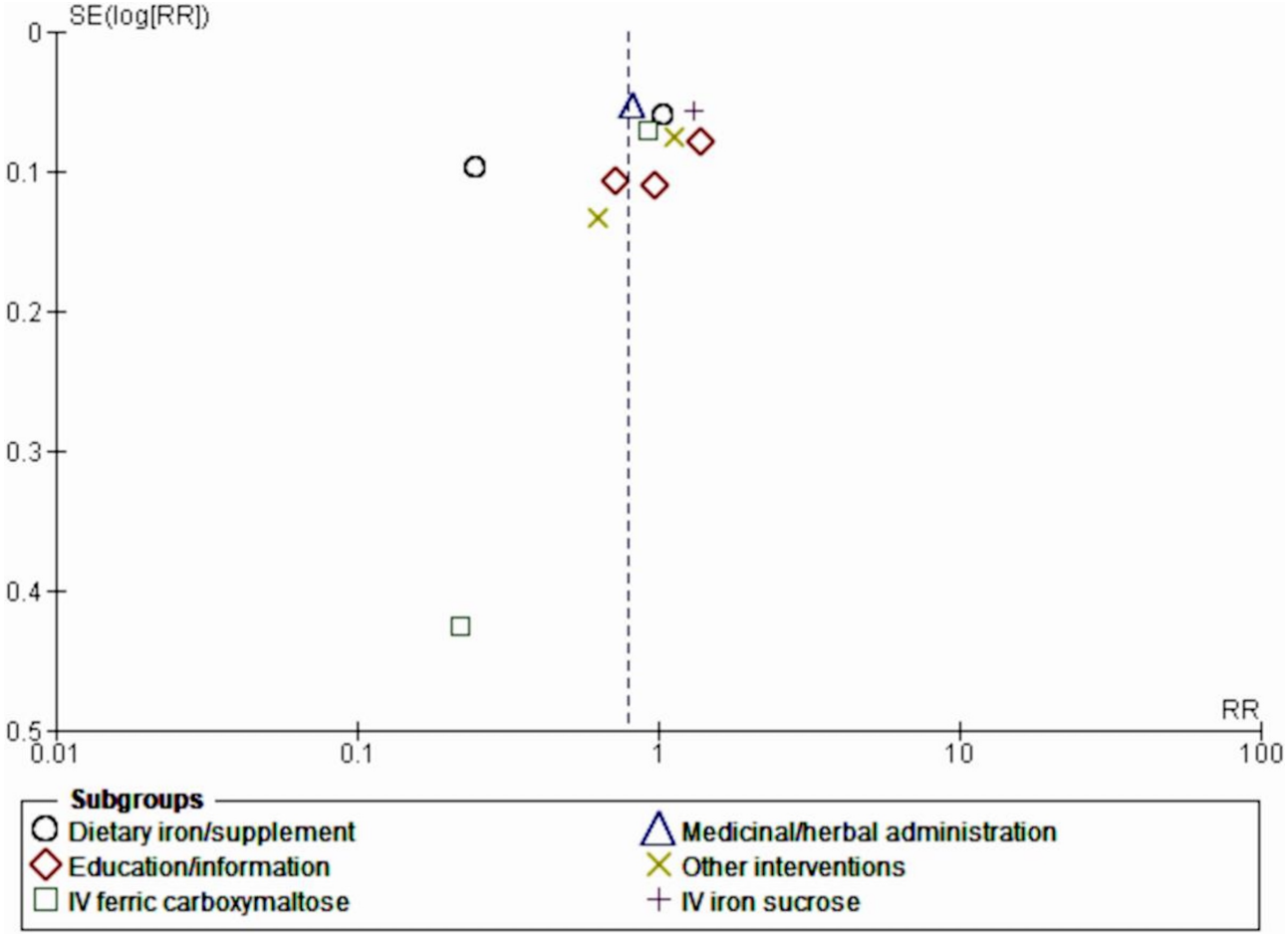
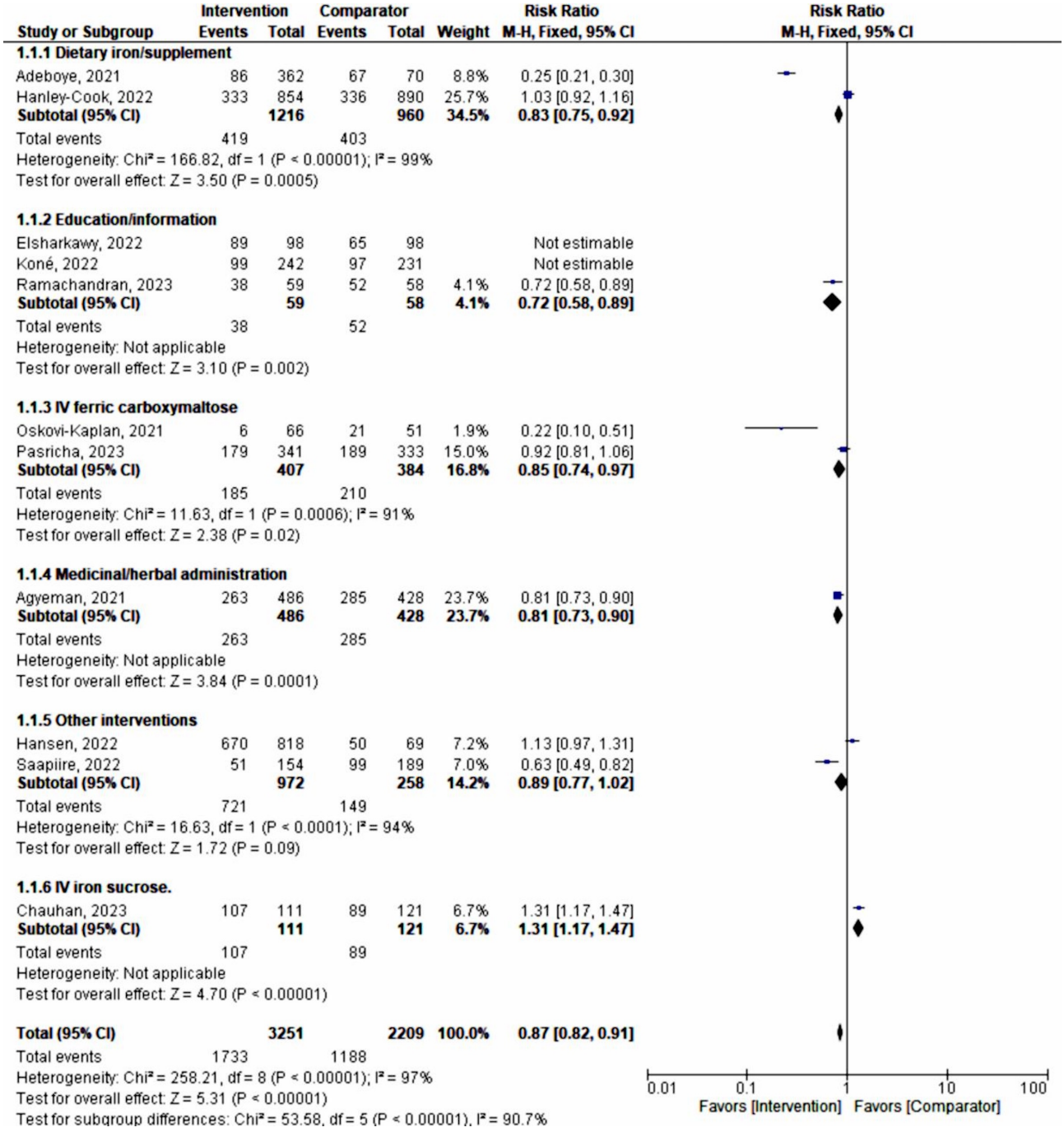


Figure 7. Funnel plot of the subgroup analysis (type or form of intervention). IV: intravenous; RR: rate ratio; SE: standard error.



Using a fixed-effect model, assuming one true effect size underlies each specific intervention form or approach, the subgroup analysis demonstrated the following significant influences on the prevention and management of maternal anemia: dietary iron supplementation (RR 0.83, 95% CI 0.75 - 0.92; *P*<.001), IV ferric carboxylmaltose (RR 0.85, 95% CI 0.74 - 0.97; *P*<.02), and medicinal or herbal administration (RR 0.81, 95% CI 0.73 - 0.90; *P*<.001). However, all interventions still exhibited high heterogeneity (I^2>90%; Figure 8).

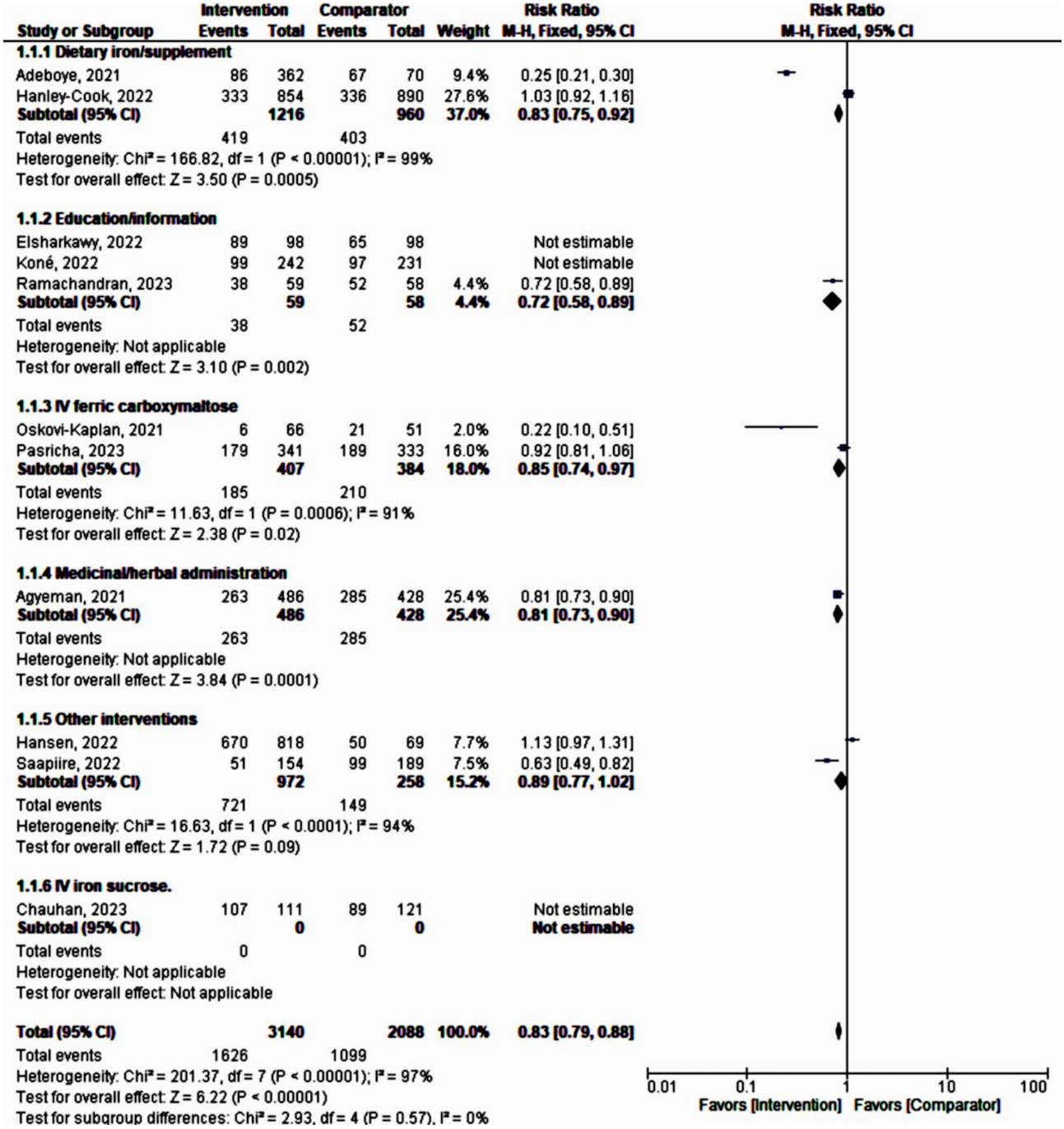
Figure 8. Subgroup analysis (fixed-effect model) according to the type or form of intervention [33-43]. M-H: Mantel-Haenszel.



The high heterogeneity obtained prompted a further sensitivity analysis on each subgroup to identify the group most strongly associated with heterogeneity. Following this analysis on subgroups (n=5228), by eliminating studies causing major heterogeneity [37,38,40], all intervention approaches against maternal anemia showed a pooled positive effect of 17% (fixed-effect model RR 0.83, 95% CI 0.79 - 0.88; *P*<.001; χ^2_4 =2.93, *P*=.57; I^2=0%).

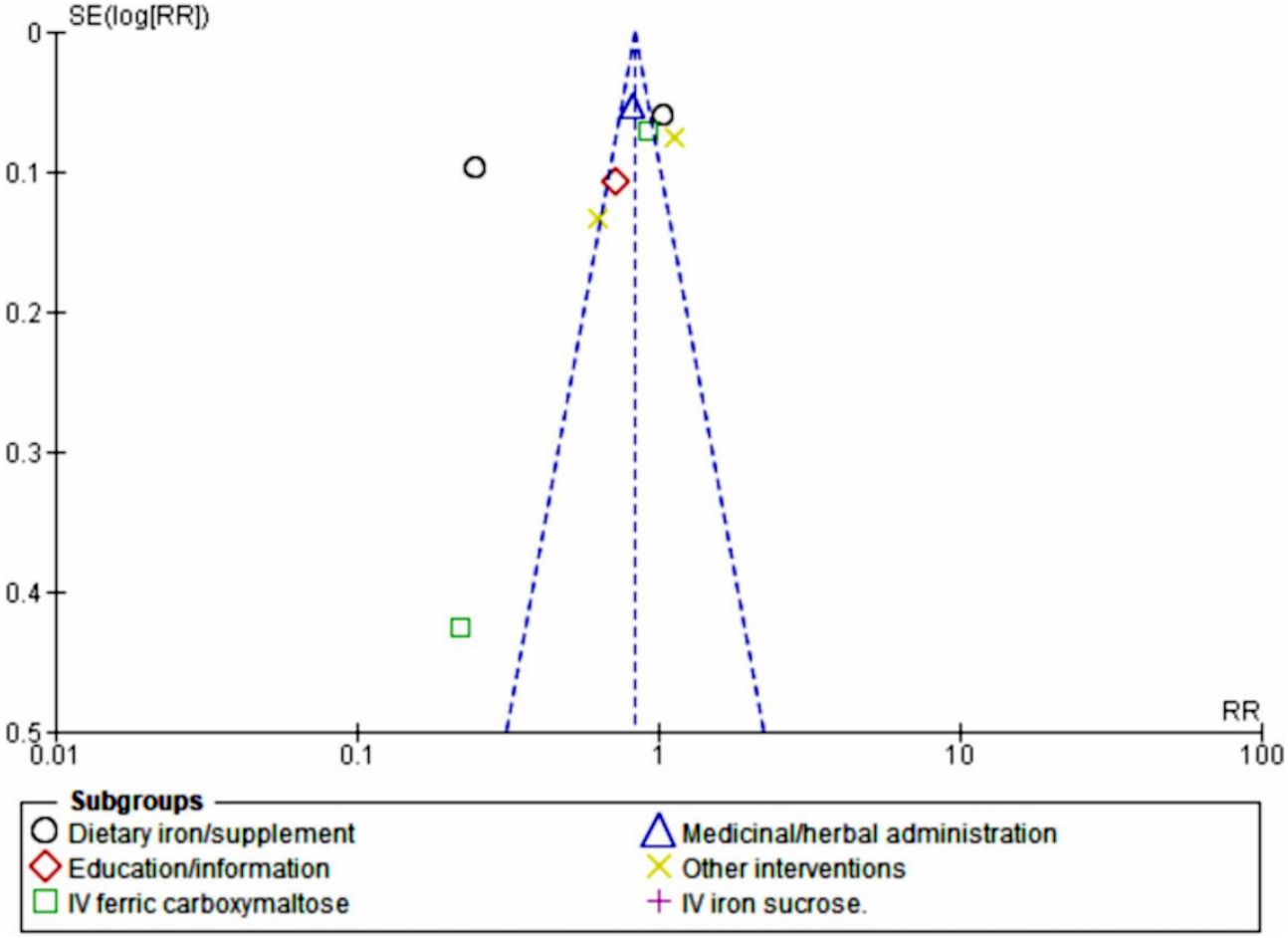
Education or information given to pregnant women (n=117) showed a 28% effect (RR 0.72, 95% CI 0.58 - 0.89; *P*<.001). Medicinal or herbal administration had a 19% effect (RR 0.81, 95% CI 0.73 - 0.90; *P*<.001; n=914). Dietary iron supplementation showed a 17% effect (RR 0.83, 95% CI 0.75 - 0.92; *P*<.001; n=2176). IV ferric carboxylmaltose exhibited a 15% effect (RR 0.85, 95% CI 0.74 - 0.97; *P*<.02; n=791; Figure 9).

Figure 9. Sensitivity analysis on intervention type subgroups showing very low heterogeneity [33-43]. M-H: Mantel-Haenszel.



These findings were accompanied by greatly reduced publication bias and heterogeneity between the subgroups (I^2=0%), as shown by the funnel plot (Figure 10).

Figure 10. Funnel plot of the subgroup sensitivity analysis (type of intervention). IV: intravenous; RR: rate ratio; SE: standard error.

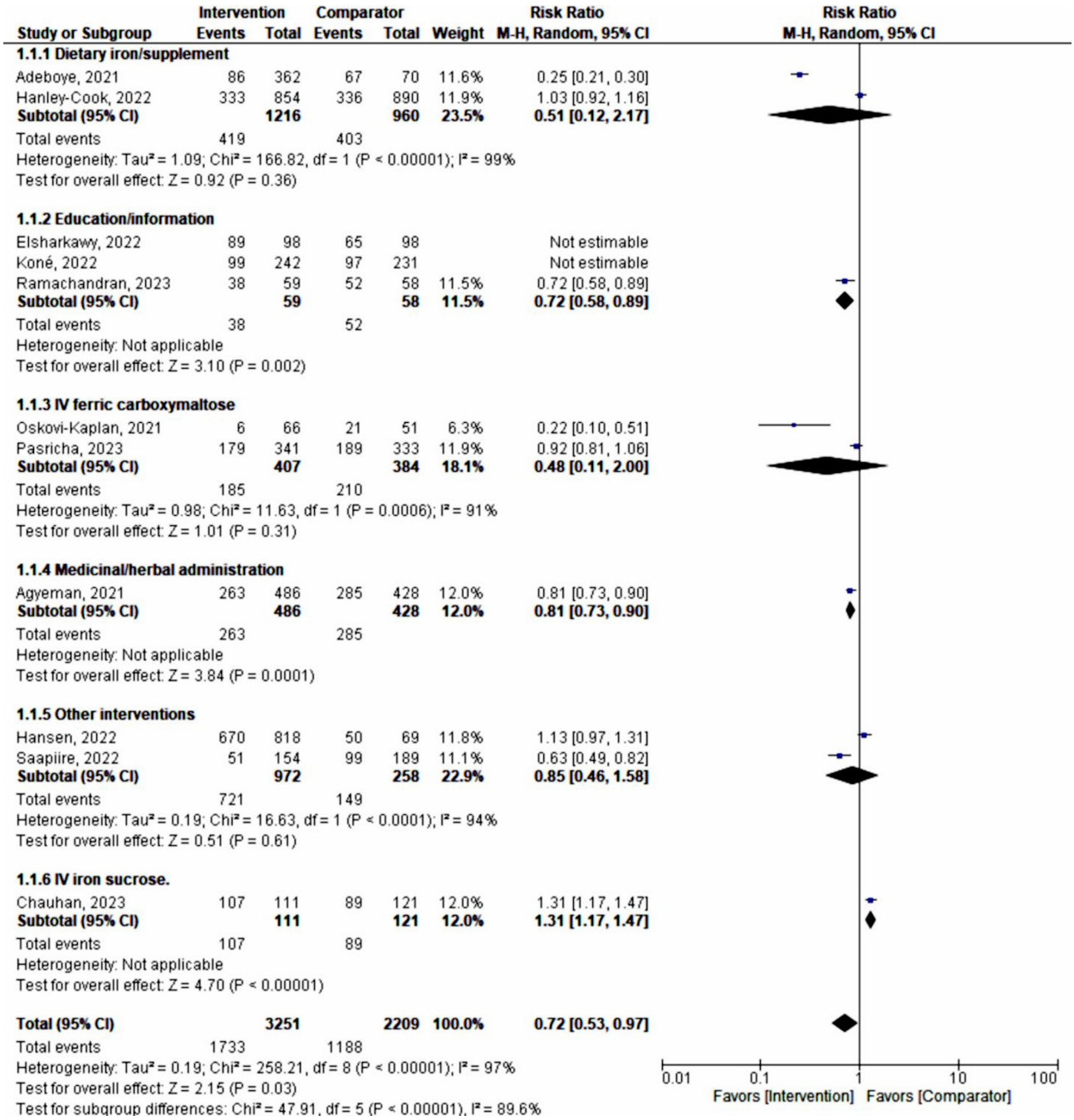


Subgroup and Sensitivity Analysis on the Possible Covariates

Location of the Pregnant Women

Generally, maternal anemia interventions during the advent of the COVID-19 pandemic demonstrated a higher and significant effect (16%) in Africa (n=4580) compared to Asia and Europe (fixed-effects model RR 0.84, 95% CI 0.79 - 0.89; P<.001; χ^2 =176.53, P<.001; I^2=97%; Figure 11).

Figure 11. Subgroup analysis by the location of pregnant women during the advent of the COVID-19 pandemic [33-43]. M-H: Mantel-Haenszel.



Study Setting

Similarly, multicenter studies (n=4580) showed a more significant predictive effect (16%) on maternal anemia intervention compared to single-center studies (n=1549; fixed-effects model RR 0.84, 95% CI 0.79 - 0.89; *P*<.001; χ^2 =176.53, *P*<.001; I^2=97%; Figure 12). The funnel plot demonstrated that most studies close to the mean effect were multicenter and associated with heterogeneity, with only one study tending to signify homogeneity (Figure 13).

Figure 12. Subgroup analysis by study setting [33-43]. M-H: Mantel-Haenszel.

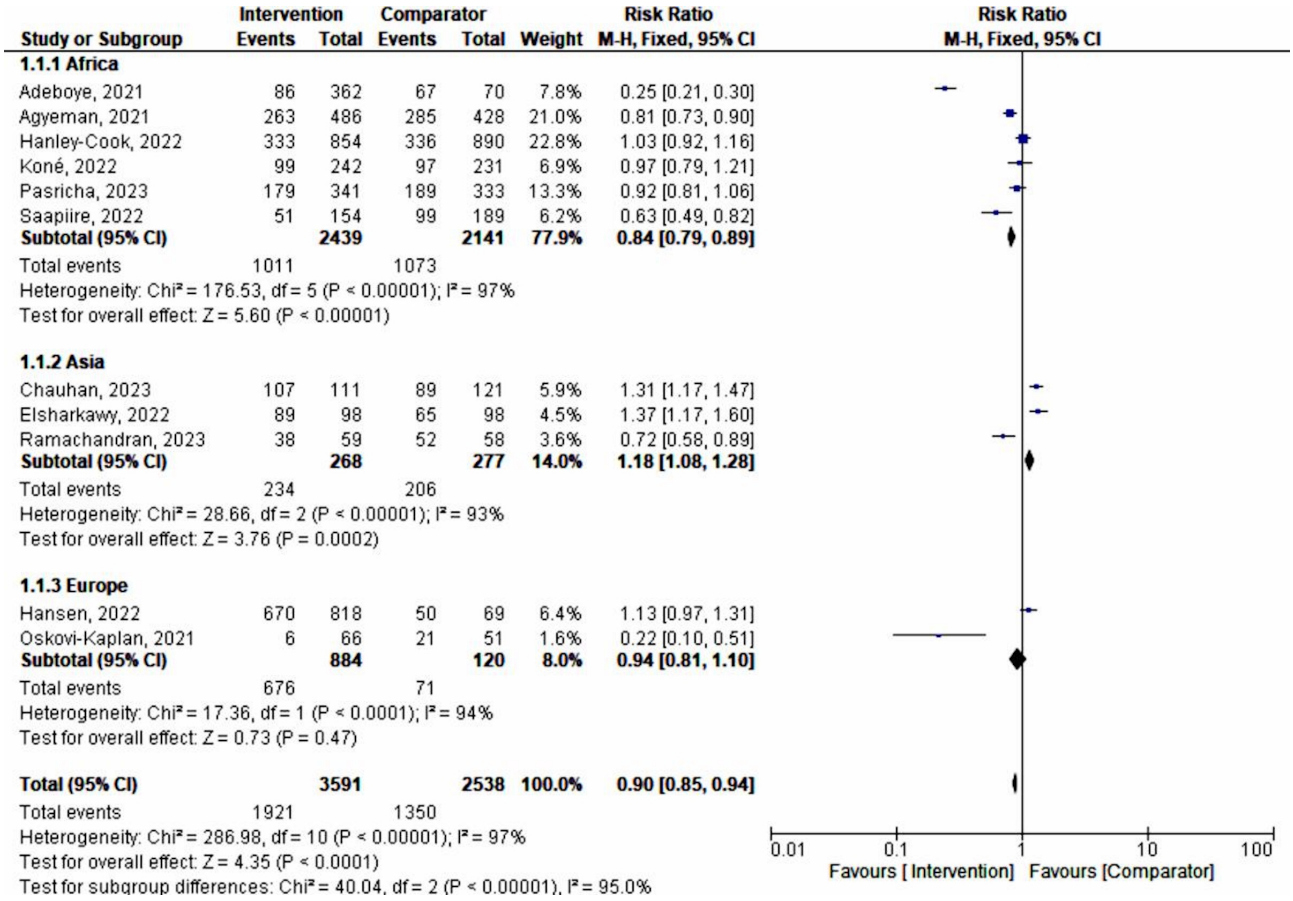
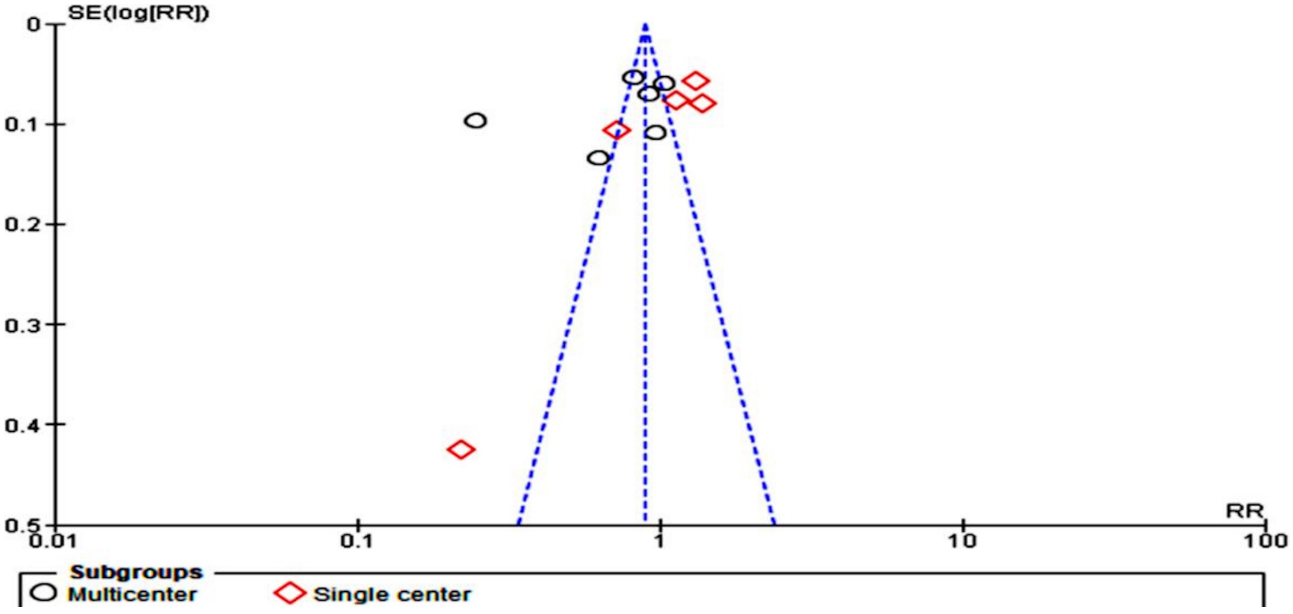


Figure 13. A funnel plot of subgroup analysis by study setting, showing that single-center studies (diamond shape) signified higher heterogeneity. RR: rate ratio; SE: standard error.



Time or Year of Data Collection

Studies whose data were collected in 2020, during the advent of the COVID-19 pandemic (n=2350), showed a significantly higher predictive effect (50%) on maternal anemia intervention compared to data from other times or years (random-effects model RR 0.50, 95% CI 0.26 - 0.99; P<.05; χ^2_3 =167.34, P<.001; I^2=98%; Figure 14). This finding was further supported by fixed-effect analysis, where the year 2020 showed a 28% effect (RR 0.72, 95% CI 0.67 - 0.78; P<.001; χ^2_3 =167.34, P<.001; I^2=98%; Figure 15).

Figure 14. Subgroup analysis by the year the data were collected (random-effects model) [33-43]. M-H: Mantel-Haenszel.

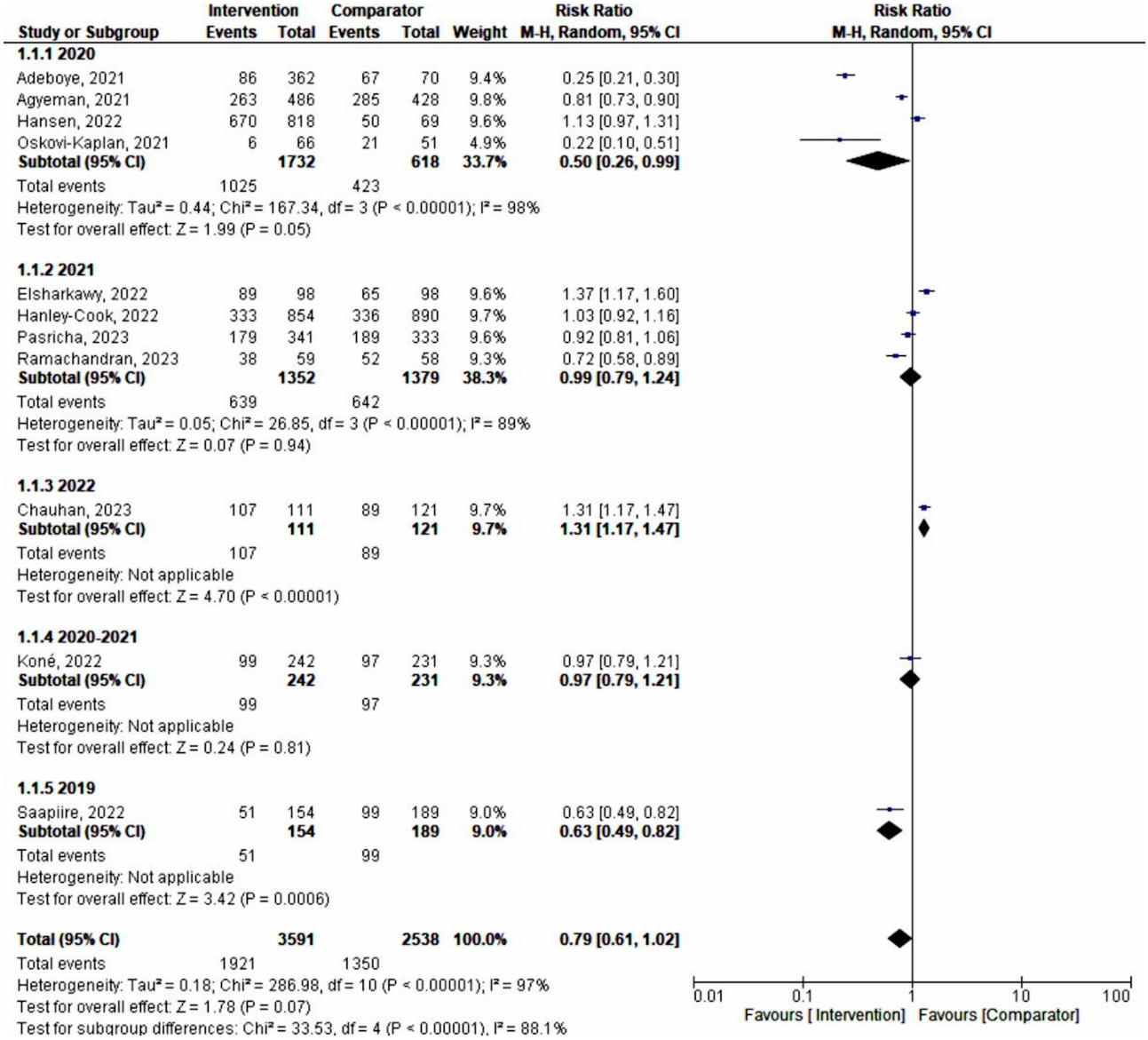
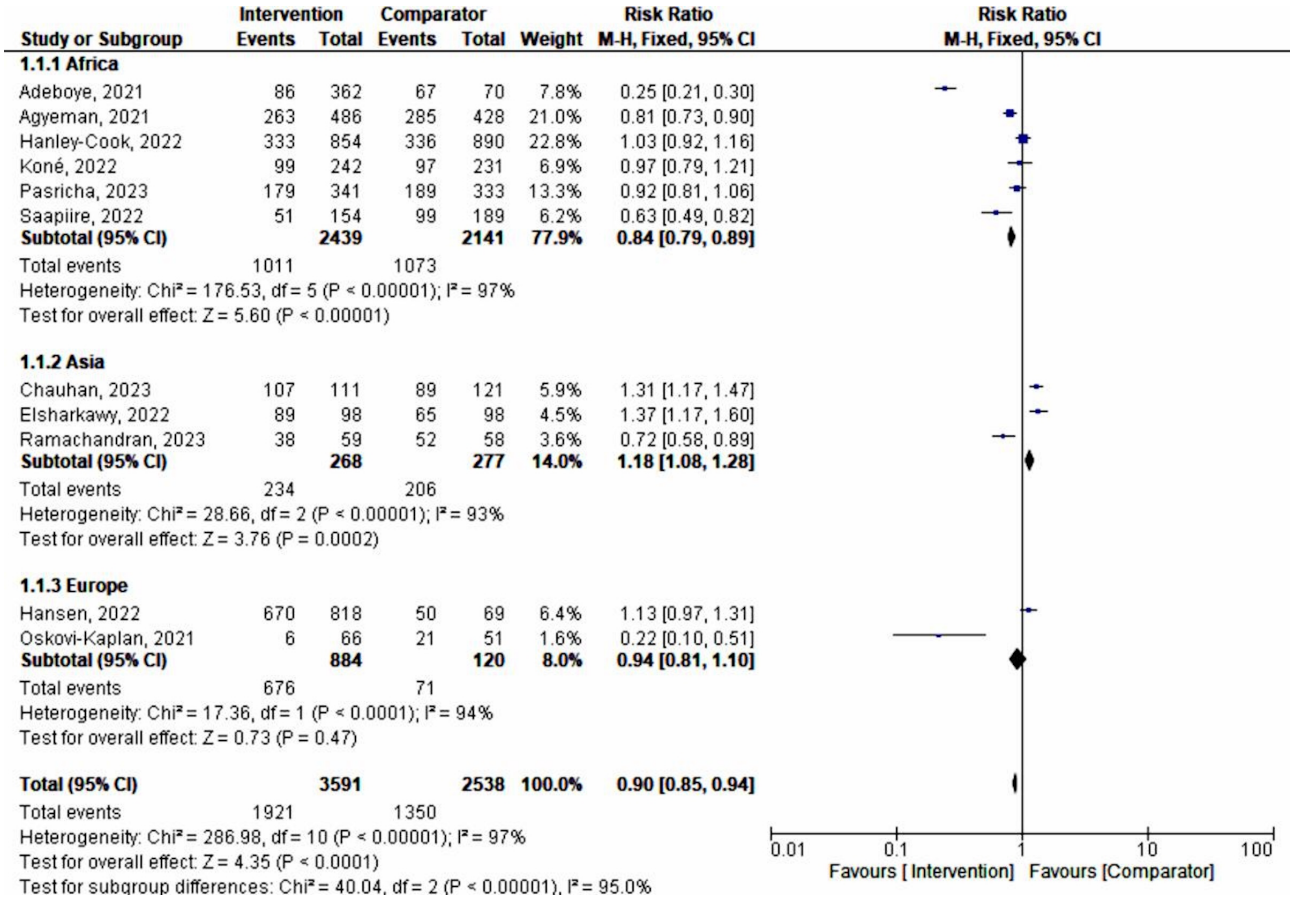


Figure 15. Subgroup analysis by the year the data were collected (fixed-effect model) 33-43. M-H: Mantel-Haenszel.



Statistical Justifications of the Use of Fixed-Effect or Random-Effects Models

The initial analysis used a random-effects model due to significant heterogeneity among the studies (I^2=97%; P<.001), allowing for variations in true effect sizes across different populations and contexts. This model was suitable for handling diverse study designs, sample sizes, and intervention approaches. However, when assuming one true effect size for specific intervention forms, a fixed-effects model was used. This model was appropriate for evaluating interventions like dietary iron supplementation, IV ferric carboxylmaltose, and medicinal or herbal administration, despite high heterogeneity (I^2>90%). By combining both models, the study leveraged the strengths of each approach, ensuring that the analysis accounted for both within-study and between-study variability. This dual approach enhanced the robustness and credibility of the findings, leading to more accurate and reliable pooled estimates for maternal anemia interventions.

Discussion

Principal Findings

This study found that while the overall effect of interventions on maternal anemia during COVID-19 was initially unclear due to high variability and publication bias, more focused analyses showed certain interventions—such as dietary iron supplementation, herbal/medicinal treatments, and IV ferric carboxylmaltose—were significantly beneficial. Sensitivity and subgroup analyses revealed these effects were strongest in studies conducted in Africa, studies conducted in multicenter settings, and especially among data collected in 2020, with a notable reduction in heterogeneity and a clearer positive impact on maternal hemoglobin levels.

The meta-analysis explored the effects of various interventions on maternal anemia among 6129 pregnant women during the COVID-19 pandemic. The primary outcome focused on the effectiveness of different interventions, such as dietary iron supplementation, education, intravenous iron therapy, and medicinal or herbal treatments. Although the overall quality of studies was moderate, the interventions showed varied effectiveness in managing and preventing maternal anemia. Sensitivity and subgroup analyses helped identify factors contributing to the observed heterogeneity in results. The interventions were generally more effective in Africa compared to Asia and Europe, and data collected during 2020 indicated a more significant impact. Overall, the study highlighted the need for further research to draw more definitive conclusions about the effectiveness of these interventions on maternal anemia.

This meta-analysis included 11 studies and revealed that the pooled intervention approaches had an effect on mitigating or reducing maternal anemia in the advent of the COVID-19 pandemic by 39%. Previous research has indicated a similar range of effect during the pandemic based on iron supplementation and iron and folic acid interventions (1.39, 0.33-2.45; P=.01 and 0.72, 0.36-1.07, P<.001, respectively) [15,27,44]. The general net cumulative effect of the interventions

on maternal anemia ranged from 23% to 81% [6,45-47]. This is supported by recently published studies. Additionally, other meta-analyses have reported the outcomes of interventions on maternal anemia during the COVID-19 pandemic [28,44]. This analysis adds to the extensive consensus in the literature, which should motivate further research investigating the key aspects inherent to anemia control in pregnancy during similar pandemics.

Further, this systematic meta-analysis offers a more detailed view as it covers 11 studies from diverse regions capturing both single-center and multicenter studies. The heterogeneity was high even after subgroup analysis adjustments as per the specific cluster of intervention.

However, with fixed model analysis, dietary iron supplementation (17%), IV ferric carboxylmaltose (15%), and medicinal or herbal administration (19%) interventions significantly influenced the prevention and or management of maternal anemia. These findings are consistent with the guidance provided in the e-Library of Evidence for Nutrition Actions, where it is noted that daily iron and folic acid supplementation during pregnancy improve anemia [48], while the efficacy of intravenous ferric carboxylmaltose has been shown to be similarly positive for the condition [27,28,49-53]. In addition, past studies have also shown similar trends of significant control of maternal anemia through the use of medicinal or herbal treatments [29,54-56]. It is important to note that the effect found in this study as demonstrated by the pooled and specific interventions is seemingly lower as compared to the effects demonstrated by the earlier studies mentioned herein. Therefore, this can possibly be attributed to the influence of the COVID-19 pandemic in compromising the effectiveness of different anemia interventions.

The sensitivity analysis with the fixed-effect model on the subgroups by intervention type further showed a pooled positive effect of 17%. Notably, education or information interventions showed a 28% effect. Medicinal or herbal administration, iron supplementation, and IV ferric carboxylmaltose also had effects. The greatly reduced publication bias and heterogeneity between the subgroups following this sensitivity analysis provides evidence that using education and information to control maternal anemia is generally an efficient approach.

Anemia intervention in Africa generally had the highest effect as compared to other regions. However, this may not be due to best practices, as the prevalence of maternal anemia is higher in sub-Saharan Africa than in other regions [4,29,57-59]; instead, it may be due to more interventions being implemented to control anemia in Africa. Data collected from multicenter studies showed a more predictive effect (16%) of maternal anemia intervention as compared to single-center studies. Similar findings were reported by other similar reviews, although not during the COVID-19 pandemic [12,44,60]. In this context, the single-center studies had major heterogeneity as compared to multicenter studies.

Studies whose data were collected in the year 2020 in the advent of the COVID-19 pandemic had a more significant predictive effect (50%) on maternal anemia intervention as compared to other times or years of data collection. A subanalysis showed that the trend in maternal anemia effectiveness decreased with time from the year 2020 to 2021 and 2022. This fact is supported by a report asserting that the availability of nutritious foods in particular was affected by COVID-19 measures [7,61]. This was expected as nations concentrated on COVID-19 mitigation when it became a pandemic 2020 onward.

In addition, micronutrient intervention programs were affected during COVID-19, including disruptions of up to 75% for antenatal care programs in selected countries during the first months of the lockdown [29,62]. Furthermore, stockouts of iron and folic acid/multiple micronutrient supplementation may have occurred due to supply chain disruptions and programs no longer reporting stock information [9,63].

Comparison to Prior Work

Prior studies have reported results that contrast with those presented here, with a better effect based on percentage reduction and/or hemoglobin mean standard deviation change on controlling maternal anemia [29,64-66]; however, these studies did not include data from the advent of the COVID-19 pandemic. In addition, a meta-analysis that targeted only RCTs [7,67] showed superiority in preventing anemia by the intervention as compared with the control. Moreover, a study focusing on hemoglobin mean level change demonstrated a similar trend in improving anemia control in pregnancy [9,68]. Of concern, as mentioned previously, most studies showed mixed outcomes relative to the outcome measure, with timelines of data collection in some being outside the scope of this study, which focused on the advent of the COVID-19 pandemic.

Similar findings were reported in a previous study, in which individual education through a pictorial handbook on anemia in conjunction with a counseling intervention program had a positive impact on hemoglobin and hematocrit levels for pregnant women with anemia in their third trimester of pregnancy [9,69]. The mean change in hemoglobin levels was also found to be significant in another study, which established that educational interventions can increase family support for maternal behaviors that can prevent anemia during pregnancy, such as improving adherence to taking iron supplements and maintaining a high intake of food containing iron [7,65]. Prior studies reported better outcomes from information package interventions compared to the current findings in the advent of the COVID-19 pandemic. Generally, an education package on maternal anemia control is part of an integrated approach where all the other intervention methods are included as part of the package [9,70]. This may be why the current findings show that this intervention had the highest effect on maternal anemia control.

The effects and impacts of specific disasters and/or calamities on maternal anemia interventions have been investigated previously. In one study, the COVID-19 crisis exacerbated maternal and child undernutrition and child mortality in low- and middle-income countries [28,71]. Further, measuring the effects of COVID-19 disruptions on the delivery of essential health and nutrition interventions has proven challenging, as resilient, real-time information systems were not well-established in many countries before the crisis [9,72]. A World Health Organization report surveyed the extent of

disruptions across all health care services; such disruptions may have included disruptions to pregnancy anemia management and interventions [73]. Similarly, a study based in Africa found that health care services utilization in the advent of the COVID-19 pandemic was disrupted [29,74]. This could be expected to have affected and compromised standard interventions for mitigating maternal health issues. Another study demonstrated that the COVID-19 pandemic affected maternal health both directly and indirectly, including poor birth and maternal health outcomes [10,75]. This can explain the reduced effect of maternal anemia interventions.

Research shows that, in 2019, global anemia prevalence was 29.9% (95% uncertainty interval [UI] 27.0%-32.8%) in women of reproductive age, equivalent to over half a billion women aged 15‐49 years. Prevalence was 29.6% (95% UI 26.6%-32.5%) in nonpregnant women of reproductive age and 36.5% (95% UI 34.0%-39.1%) in pregnant women. Since 2000, the global prevalence of anemia in women of reproductive age has been stagnant, while the prevalence of anemia in pregnant women has decreased slightly [27,76]. Although more information is accumulating daily since the COVID-19 pandemic, subjective factors on pregnancy and the effect of the pandemic on health systems in African nations may have compromised the progress toward addressing anemia in general [29,77]. Given this, a few interlinked factors, including any similar pandemics, should be considered together as a single risk factor for maternal anemia.

Strengths and Limitations

The study included 11 articles with 6129 participants and revealed a pooled intervention effect of 39% in preventing and managing maternal anemia. Interventions such as education (28%), medicinal administration (19%), iron supplementation (17%), and IV ferric carboxylmaltose (15%) showed substantial impact, especially in Africa. Multicenter studies were more predictive than single-center ones. Sensitivity analyses significantly reduced heterogeneity (I^2=0%), increasing the reliability of results. Education and tailored strategies proved highly effective in low-resource settings and during crises, highlighting the importance of contextual interventions.

Several constraints may have affected the findings. First, most studies were retrospective, with only 4 RCTs contributing high-quality evidence. This could weaken the robustness of pooled estimates, though sensitivity analyses were conducted to mitigate this. Second, a lack of demographic details—such as participant age or gestational stage—led to inconsistent data and reduced comparability. Future studies should ensure thorough reporting to enhance clarity. Third, COVID-19 may have indirectly affected anemia metrics through influences on hemoglobin levels and nutritional access. These impacts underscore the need for pandemic-adjusted assessments in future meta-analyses. Publication bias may have exaggerated effectiveness, but comprehensive searches and statistical adjustments helped maintain validity.

Future Directions

The effectiveness of maternal anemia interventions declined during the COVID-19 pandemic (2020‐2022), even for the most reliable approaches. Future pandemics call for rapid research into resilient solutions. Pregnant women should be screened for tailored interventions, and stakeholders must prioritize maternal health in emergency planning. Further studies should explore the mechanisms behind the reduced effectiveness and improve delivery systems.

This meta-analysis advances existing knowledge by using rigorous methodologies and expanded datasets from 11 articles with 6129 participants. It reveals novel insights, such as a 39% utility in preventing and managing maternal anemia, with significant impacts from education (28%), medicinal administration (19%), iron supplementation (17%), and IV ferric carboxylmaltose (15%). The study also highlights regional differences, particularly higher effectiveness in Africa, and underscores the importance of multicenter studies and ongoing research.

Feasible Policy Recommendations

Based on this study, we make the following policy recommendations:

- Tailored regional interventions: focus on region-specific approaches, especially in Africa, to address unique challenges and maximize effectiveness.
- High-impact interventions: emphasize proven interventions like dietary iron supplementation, intravenous ferric carboxylmaltose, and medicinal or herbal administration for better prevention and management.
- Multicenter studies: encourage multicenter studies to improve the generalizability and reliability of results.
- Education and information: educate pregnant women about anemia prevention and management, providing dietary information and education.
- Time-specific considerations: tailor interventions to address unique circumstances, such as the advent of future pandemics, and resulting impacts on maternal anemia.

Conclusion

The COVID-19 pandemic exposed critical gaps in maternal anemia management, underscoring the need for resilient health care strategies and enhanced data systems. Although meta-analytical evidence revealed modest yet significant intervention effects—especially from medicinal or herbal therapies, education, and dietary iron supplementation—these benefits were most evident in multicenter studies and African populations when high-heterogeneity outliers were excluded. This context-specific efficacy highlights the urgency for tailored approaches and further research to strengthen maternal and child health during future global crises, as emphasized [10,78].

Acknowledgments

This study was funded partially by Kenya Medical Training College (grant number 2023-2-001). The funder had no role in the study design or interpretation.

Data Availability

All data generated or analyzed during this study are included in this published article and its supplementary files. Additional data may be requested from the corresponding author.

Authors' Contributions

Conceptualization: JKM and DKMF.
Methodology: JKM and FMW.
Data curation: JKM and KO.
Formal analysis: JKM.
Writing – original draft: JKM.
Writing – review & editing: all authors.
Supervision: RN.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Supplementary tables.
[DOCX File, 22 KB - xmed_v6i1e57626_app1.docx]

Checklist 1
PRISMA checklist.
[DOCX File, 30 KB - xmed_v6i1e57626_app2.docx]

References

1. Moghaddam Tabrizi F, Barjasteh S. Maternal hemoglobin levels during pregnancy and their association with birth weight of neonates. Iran J Ped Hematol Oncol 2015;5(4):211-217. [Medline: 26985354]
2. Chaparro CM, Suchdev PS. Anemia epidemiology, pathophysiology, and etiology in low- and middle-income countries. Ann N Y Acad Sci 2019 Aug;1450(1):15-31. [doi: 10.1111/nyas.14092] [Medline: 31008520]
3. Yang J, Liu Z, Guo H, et al. Prevalence and influencing factors of anaemia among pregnant women in rural areas of Northwestern China. Public Health (Fairfax) 2023 Jul;220:50-56. [doi: 10.1016/j.puhe.2023.04.024] [Medline: 37269588]
4. Anaemia in women and children: WHO global anaemia estimates. World Health Organization. 2021. URL: https://www.who.int/data/gho/data/themes/topics/anaemia_in_women_and_children [accessed 2025-09-23]
5. Rocca-Ihenacho L, Alonso C. Where do women birth during a pandemic? Changing perspectives on Safe Motherhood during the COVID-19 pandemic. J Glob Health Sci 2020;2(1). [doi: 10.35500/jghs.2020.2.e4]
6. Benson AE, Martens KL, Ryan KS, et al. Correction of anemia with intravenous iron mitigates adverse maternal outcomes in pregnancy. Blood 2023 Nov 2;142(Supplement 1):3747-3747. [doi: 10.1182/blood-2023-178359]
7. Flaherty SJ, Delaney H, Matvienko-Sikar K, Smith V. Maternity care during COVID-19: a qualitative evidence synthesis of women's and maternity care providers' views and experiences. BMC Pregnancy Childbirth 2022 May 26;22(1):438. [doi: 10.1186/s12884-022-04724-w] [Medline: 35619069]
8. Kumar N. COVID 19 era: a beginning of upsurge in unwanted pregnancies, unmet need for contraception and other women related issues. Eur J Contracept Reprod Health Care 2020 Aug;25(4):323-325. [doi: 10.1080/13625187.2020.1777398] [Medline: 32567961]
9. Schmitt N, Mattern E, Cignacco E, et al. Effects of the Covid-19 pandemic on maternity staff in 2020 - a scoping review. BMC Health Serv Res 2021 Dec 27;21(1):1364. [doi: 10.1186/s12913-021-07377-1] [Medline: 34961510]
10. Kotlar B, Gerson EM, Petrillo S, Langer A, Tiemeier H. Correction: The impact of the COVID-19 pandemic on maternal and perinatal health: a scoping review. Reprod Health 2023 Mar 30;20(1):52. [doi: 10.1186/s12978-023-01575-2] [Medline: 36998017]
11. Heidkamp R, Guida R, Phillips E, Clermont A. The Lives Saved Tool (LiST) as a model for prevention of anemia in women of reproductive age. J Nutr 2017 Nov;147(11):2156S-2162S. [doi: 10.3945/jn.117.252429] [Medline: 28904114]

12. Skolmowska D, Głąbska D, Kołota A, Guzek D. Effectiveness of dietary interventions in prevention and treatment of iron-deficiency anemia in pregnant women: a systematic review of randomized controlled trials. Nutrients 2022 Jul 23;14(15):3023. [doi: 10.3390/nu14153023] [Medline: 35893877]
13. Skolmowska D, Głąbska D, Kołota A, Guzek D. Effectiveness of dietary interventions to treat iron-deficiency anemia in women: a systematic review of randomized controlled trials. Nutrients 2022 Jun 30;14(13):2724. [doi: 10.3390/nu14132724] [Medline: 35807904]
14. Maddock J, Parsons S, Di Gessa G, et al. Inequalities in healthcare disruptions during the COVID-19 pandemic: evidence from 12 UK population-based longitudinal studies. BMJ Open 2022 Oct 13;12(10):e064981. [doi: 10.1136/bmjopen-2022-064981] [Medline: 36229151]
15. Yang X, Yu Y, Xu J, et al. Clinical course and outcomes of critically ill patients with SARS-CoV-2 pneumonia in Wuhan, China: a single-centered, retrospective, observational study. Lancet Respir Med 2020 May;8(5):475-481. [doi: 10.1016/S2213-2600(20)30079-5] [Medline: 32105632]
16. Singh SS, Singh LB. Training community health workers for the COVID-19 response, India. Bull World Health Organ 2022 Feb 1;100(2):108-114. [doi: 10.2471/BLT.21.286902] [Medline: 35125535]
17. Tan SY, Foo CD, Verma M, et al. Mitigating the impacts of the COVID-19 pandemic on vulnerable populations: lessons for improving health and social equity. Soc Sci Med 2023 Jul;328:116007. [doi: 10.1016/j.socscimed.2023.116007] [Medline: 37279639]
18. da Silva Lopes K, Yamaji N, Rahman MO, et al. Nutrition-specific interventions for preventing and controlling anaemia throughout the life cycle: an overview of systematic reviews. Cochrane Database Syst Rev 2021 Sep 26;9(9):CD013092. [doi: 10.1002/14651858.CD013092.pub2] [Medline: 34564844]
19. Perelman SI, Shander A, Mabry C, Ferraris VA. Preoperative anemia management in the coronavirus disease (COVID-19) era. JTCVS Open 2021 Mar;5:85-94. [doi: 10.1016/j.xjon.2020.12.020] [Medline: 34173552]
20. e-Library of Evidence for Nutrition Actions (eLENA). Exclusive breastfeeding for optimal growth, development and health of infants. World Health Organization. 2023. URL: https://www.who.int/tools/elena/interventions/exclusive-breastfeeding [accessed 2025-09-23]
21. Jin Q, Shimizu M, Sugiura M, et al. Effectiveness of non-pharmacological interventions to prevent anemia in pregnant women: a quantitative systematic review protocol. JBI Evid Synth 2024 Jun 1;22(6):1122-1128. [doi: 10.11124/JBIES-23-00081] [Medline: 38084098]
22. Global nutrition targets 2025: anaemia policy brief. World Health Organization. 2014. URL: https://www.who.int/publications/i/item/WHO-NMH-NHD-14.4 [accessed 2025-09-23]
23. Jamison DT, Alwan A, Mock CN. Universal health coverage and intersectoral action for health: key messages from Disease Control Priorities, 3rd edition. The Lancet 2018 Mar;391(10125):1108-1120. [doi: 10.1016/S0140-6736(15)60097-6]
24. Brandibur TE, Kundnani NR, Boia M, et al. Does COVID-19 infection during pregnancy increase the appearance of congenital gastrointestinal malformations in neonates? Biomedicines 2023 Nov 21;11(12):3105. [doi: 10.3390/biomedicines11123105] [Medline: 38137326]
25. Jamieson DJ, Rasmussen SA. An update on COVID-19 and pregnancy. Am J Obstet Gynecol 2022 Feb;226(2):177-186. [doi: 10.1016/j.ajog.2021.08.054] [Medline: 34534497]
26. Khedmat L, Mohaghegh P, Veysizadeh M, Hosseinkhani A, Fayazi S, Mirzadeh M. Pregnant women and infants against the infection risk of COVID-19: a review of prenatal and postnatal symptoms, clinical diagnosis, adverse maternal and neonatal outcomes, and available treatments. Arch Gynecol Obstet 2022 Aug;306(2):323-335. [doi: 10.1007/s00404-021-06325-y] [Medline: 34842975]
27. Zhang J, Li Q, Song Y, Fang L, Huang L, Sun Y. Nutritional factors for anemia in pregnancy: a systematic review with meta-analysis. Front Public Health 2022;10. [doi: 10.3389/fpubh.2022.1041136]
28. . R, Widowati R, Siauta JA, Azzahroh P, Silawati V. The prevention of anaemia among pregnant women: a literature review. ijmst 2023;10(2):2867-2872. [doi: 10.15379/ijmst.v10i2.2978]
29. Rahma FA. Anaemia in pregnancy: a literature review. J Ilm STIKES Yars Mataram 2023;13(1). [doi: 10.57267/jisym.v13i1.233]
30. Moher D, Shamseer L, Clarke M, et al. Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015 statement. Syst Rev 2015 Jan 1;4(1):1. [doi: 10.1186/2046-4053-4-1] [Medline: 25554246]
31. Jørgensen L, Paludan-Müller AS, Laursen DRT, et al. Evaluation of the Cochrane tool for assessing risk of bias in randomized clinical trials: overview of published comments and analysis of user practice in Cochrane and non-Cochrane reviews. Syst Rev 2016 May 10;5(1):80. [doi: 10.1186/s13643-016-0259-8] [Medline: 27160280]
32. Zeng X, Zhang Y, Kwong JSW, et al. The methodological quality assessment tools for preclinical and clinical studies, systematic review and meta‐analysis, and clinical practice guideline: a systematic review. J Evidence Based Medicine 2015 Feb;8(1):2-10. [doi: 10.1111/jebm.12141]
33. Oskovi-Kaplan ZA, Kilickiran H, Buyuk GN, Ozyer S, Keskin HL, Engin-Ustun Y. Comparison of the maternal and neonatal outcomes of pregnant women whose anemia was not corrected before delivery and pregnant women who were treated with intravenous iron in the third trimester. Arch Gynecol Obstet 2021 Mar;303(3):715-719. [doi: 10.1007/s00404-020-05817-7] [Medline: 32990783]

34. Adeboye TE, Bodunde IO, Okekunle AP. Dietary iron intakes and odds of iron deficiency anaemia among pregnant women in Ifako-Ijaiye, Lagos, Nigeria: a cross-sectional study. Pan Afr Med J 2022;42:23. [doi: 10.11604/pamj.2022.42.23.29965] [Medline: 35910059]
35. Saapiire F, Dogoli R, Mahama S. Adequacy of antenatal care services utilisation and its effect on anaemia in pregnancy. J Nutr Sci 2022;11:e80. [doi: 10.1017/jns.2022.80] [Medline: 36304821]
36. Hansen R, Spangmose AL, Sommer VM, et al. Maternal first trimester iron status and its association with obstetric and perinatal outcomes. Arch Gynecol Obstet 2022 Oct;306(4):1359-1371. [doi: 10.1007/s00404-022-06401-x] [Medline: 35088196]
37. Elsharkawy NB, Abdelaziz EM, Ouda MM, Oraby FA. Effectiveness of health information package program on knowledge and compliance among pregnant women with anemia: a randomized controlled trial. Int J Environ Res Public Health 2022 Feb 26;19(5):2724. [doi: 10.3390/ijerph19052724] [Medline: 35270420]
38. Koné S, Probst-Hensch N, Dao D, Utzinger J, Fink G. Improving coverage of antenatal iron and folic acid supplementation and malaria prophylaxis through targeted information and home deliveries in Côte d'Ivoire: a cluster randomised controlled trial. BMJ Glob Health 2023 Apr;8(4):e010934. [doi: 10.1136/bmjgh-2022-010934] [Medline: 37076197]
39. Agyeman YN, Newton S, Annor RB, Owusu-Dabo E. Intermittent preventive treatment comparing two versus three doses of sulphadoxine pyrimethamine (IPTp-SP) in the prevention of anaemia in pregnancy in Ghana: a cross-sectional study. PLoS ONE 2021;16(4):e0250350. [doi: 10.1371/journal.pone.0250350] [Medline: 33878140]
40. Chauhan N, Dogra P, Sharma R, Kant S, Soni M. Randomized controlled trial comparing ferrous sulfate and iron sucrose in iron deficiency anemia in pregnancy. Cureus 2023 Feb;15(2):e34858. [doi: 10.7759/cureus.34858] [Medline: 36923182]
41. Hanley-Cook G, Toe LC, Tesfamariam K, et al. Fortified balanced energy-protein supplementation, maternal anemia, and gestational weight gain: a randomized controlled efficacy trial among pregnant women in rural Burkina Faso. J Nutr 2022 Oct 6;152(10):2277-2286. [doi: 10.1093/jn/nxac171] [Medline: 35906874]
42. Pasricha SR, Mwangi MN, Moya E, et al. Ferric carboxymaltose versus standard-of-care oral iron to treat second-trimester anaemia in Malawian pregnant women: a randomised controlled trial. The Lancet 2023 May;401(10388):1595-1609. [doi: 10.1016/S0140-6736(23)00278-7]
43. Ramachandran R, Dash M, Adaikaladorai FC, Aridass J, Zachariah B, Manoharan B. Effect of individual nutrition education on perceptions of nutritional iron supplementation, adherence to iron - folic acid intake and Hb levels among a cohort of anemic South Indian pregnant women. J Matern Fetal Neonatal Med 2023 Dec 31;36(1). [doi: 10.1080/14767058.2023.2183749]
44. Panchal PD, Ravalia A, Rana R, et al. Impact of nutrition interventions for reduction of anemia in women of reproductive age in low- and middle-income countries: a meta-review. Curr Dev Nutr 2022 Dec;6(12):nzac134. [doi: 10.1093/cdn/nzac134] [Medline: 36601436]
45. Luwangula AK, McGough L, Tetui M, et al. Improving iron and folic acid supplementation among pregnant women: an implementation science approach in East-Central Uganda. Glob Health Sci Pract 2022 Dec 21;10(6):e2100426. [doi: 10.9745/GHSP-D-21-00426] [Medline: 36951283]
46. Berhane A, Belachew T. Effect of preconception pictured-based health education and counseling on adherence to iron-folic acid supplementation to improve maternal pregnancy and birth outcome among women who plan to pregnant: "randomized control trial". Clinical Nutrition Open Science 2022 Feb;41:98-105. [doi: 10.1016/j.nutos.2021.12.002]
47. Stewart T, Lambourne J, Thorp-Jones D, Thomas DW. Implementation of early management of iron deficiency in pregnancy during the SARS-CoV-2 pandemic. Eur J Obstet Gynecol Reprod Biol 2021 Mar;258:60-62. [doi: 10.1016/j.ejogrb.2020.12.055] [Medline: 33418463]
48. Daily iron and folic acid supplementation during pregnancy. World Health Organization. 2020. URL: https://www.who.int/tools/elena/interventions/daily-iron-pregnancy [accessed 2025-09-23]
49. Kim MS, Koh IJ, Choi KY, Yang SC, In Y. Efficacy and safety of intravenous ferric carboxymaltose in patients with postoperative anemia following same-day bilateral total knee arthroplasty: a randomized controlled trial. J Clin Med 2021 Apr 2;10(7):1457. [doi: 10.3390/jcm10071457] [Medline: 33918110]
50. Froessler B, Gajic T, Dekker G, Hodyl NA. Treatment of iron deficiency and iron deficiency anemia with intravenous ferric carboxymaltose in pregnancy. Arch Gynecol Obstet 2018 Jul;298(1):75-82. [doi: 10.1007/s00404-018-4782-9] [Medline: 29740690]
51. Gandotra N, Zargar S, Mahajan N. Intravenous ferric carboxymaltose for anaemia in pregnancy. Int J Res Med Sci 2020;8(10):3539. [doi: 10.18203/2320-6012.ijrms20204225]
52. Froessler B, Collingwood J, Hodyl NA, Dekker G. Intravenous ferric carboxymaltose for anaemia in pregnancy. BMC Pregnancy Childbirth 2014 Mar 25;14(1):115. [doi: 10.1186/1471-2393-14-115] [Medline: 24667031]
53. Agrawal D, Masand DL. A study for efficacy and safety of ferric carboxymaltose versus iron sucrose in iron deficiency anemia among pregnant women in tertiary care hospital. Int J Reprod Contracept Obstet Gynecol 2019;8(6):2280. [doi: 10.18203/2320-1770.ijrcog20192418]
54. El Hajj M, Sitali DC, Vwalika B, Holst L. "Back to Eden": an explorative qualitative study on traditional medicine use during pregnancy among selected women in Lusaka Province, Zambia. Complement Ther Clin Pract 2020 Aug;40:101225. [doi: 10.1016/j.ctcp.2020.101225] [Medline: 32798811]

55. Nalumansi PA, Kamatenesi-Mugisha M, Anywar G. Medicinal plants used during antenatal care by pregnant women in Eastern Uganda. Afr J Reprod Health 2017 Dec;21(4):33-44. [doi: 10.29063/ajrh2017/v21i4.4] [Medline: 29624949]
56. Ahmed M, Hwang JH, Ali MN, Al-Ahnoumy S, Han D. Irrational use of selected herbal medicines during pregnancy: a pharmacoepidemiological evidence from Yemen. Front Pharmacol 2022;13:926449. [doi: 10.3389/fphar.2022.926449] [Medline: 35928277]
57. Abdallah F, John SE, Hancy A, et al. Prevalence and factors associated with anaemia among pregnant women attending reproductive and child health clinics in Mbeya region, Tanzania. PLOS Glob Public Health 2022;2(10):e0000280. [doi: 10.1371/journal.pgph.0000280] [Medline: 36962486]
58. Ministry of Health, Community Development, Gender, Elderly and Children, Tanzania Mainland, Ministry of Health, National Bureau of Statistics, Office of the Chief Government Statistician, ICF. 2015-16 TDHS-MIS Key Findings.: MoHCDGEC, MoH, NBS, OCGS, and ICF; 2016 URL: https://preview.dhsprogram.com/pubs/pdf/SR233/SR233.pdf [accessed 2025-09-23]
59. Odhiambo JN, Sartorius B. Mapping of anaemia prevalence among pregnant women in Kenya (2016-2019). BMC Pregnancy Childbirth 2020 Nov 23;20(1):711. [doi: 10.1186/s12884-020-03380-2] [Medline: 33228585]
60. Lopez de Romaña D, Mildon A, Golan J, Jefferds MED, Rogers LM, Arabi M. Review of intervention products for use in the prevention and control of anemia. Ann N Y Acad Sci 2023 Nov;1529(1):42-60. [doi: 10.1111/nyas.15062] [Medline: 37688369]
61. Laborde D, Martin W, Vos R. Impacts of COVID-19 on global poverty, food security, and diets: insights from global model scenario analysis. Agric Econ 2021 May;52(3):375-390. [doi: 10.1111/agec.12624] [Medline: 34230728]
62. Tracking the situation of children during COVID-19 dashboard. Standing Together for Nutrition. 2020. URL: https://knowledgehub.standingtogetherfornutrition.org/knowledge-hub/tracking-the-situation-of-children-during-covid-19-dashboard/ [accessed 2025-09-23]
63. Negro-Calduch E, Azzopardi-Muscat N, Nitzan D, Pebody R, Jorgensen P, Novillo-Ortiz D. Health information systems in the COVID-19 pandemic: a short survey of experiences and lessons learned from the European region. Front Public Health 2021;9:676838. [doi: 10.3389/fpubh.2021.676838] [Medline: 34650946]
64. Ouédraogo S, Koura GK, Bodeau-Livinec F, Accrombessi MMK, Massougbodji A, Cot M. Maternal anemia in pregnancy: assessing the effect of routine preventive measures in a malaria-endemic area. Am J Trop Med Hyg 2013 Feb;88(2):292-300. [doi: 10.4269/ajtmh.12-0195] [Medline: 23296448]
65. Triharini M, Armini NKA, Nastiti AA. Effect of educational intervention on family support for pregnant women in preventing anemia. Belitung Nurs J 2018;4(3):304-311. [doi: 10.33546/bnj.332]
66. Haider BA, Olofin I, Wang M, et al. Anaemia, prenatal iron use, and risk of adverse pregnancy outcomes: systematic review and meta-analysis. BMJ 2013 Jun 21;346(jun21 3):f3443. [doi: 10.1136/bmj.f3443] [Medline: 23794316]
67. Hansen R, Sommer VM, Pinborg A, et al. Intravenous ferric derisomaltose versus oral iron for persistent iron deficient pregnant women: a randomised controlled trial. Arch Gynecol Obstet 2023 Oct;308(4):1165-1173. [doi: 10.1007/s00404-022-06768-x] [Medline: 36107229]
68. Pinsuwan S, Chatchawet W, Chunuan S. Effectiveness of interactive learning via multimedia technology with family support program among pregnant women with anemia: a quasi-experimental study. Pacific Rim Int J Nurs Res 2022;26(4).
69. Nahrisah P, Somrongthong R, Viriyautsahakul N, Viwattanakulvanid P, Plianbangchang S. Effect of integrated pictorial handbook education and counseling on improving anemia status, knowledge, food intake, and iron tablet compliance among anemic pregnant women in Indonesia: a quasi-experimental study. J Multidiscip Healthc 2020;13:43-52. [doi: 10.2147/JMDH.S213550] [Medline: 32021233]
70. Focusing on anaemia: towards an integrated approach for effective anaemia control. World Health Organization. 2004. URL: https://www.who.int/publications/m/item/focusing-on-anaemia-towards-an-integrated-approach-for-effective-anaemia-control [accessed 2025-09-23]
71. Osendarp S, Akuoku JK, Black RE, et al. The COVID-19 crisis will exacerbate maternal and child undernutrition and child mortality in low- and middle-income countries. Nat Food 2021 Jul;2(7):476-484. [doi: 10.1038/s43016-021-00319-4] [Medline: 37117686]
72. Polašek O, Wazny K, Adeloye D, et al. Research priorities to reduce the impact of COVID-19 in low- and middle-income countries. J Glob Health 2022;12:09003. [doi: 10.7189/jogh.12.09003] [Medline: 35475006]
73. Pulse survey on continuity of essential health services during the COVID-19 pandemic: interim report. World Health Organization. 2020 Aug 27. URL: https://www.who.int/teams/integrated-health-services/health-services-performance-assessment/monitoring-health-services/global-pulse-survey-on-continuity-of-essential-health-services-during-the-covid-19-pandemic [accessed 2025-09-23]
74. Mebratie AD, Nega A, Gage A, Mariam DH, Eshetu MK, Arsenault C. Effect of the COVID-19 pandemic on health service utilization across regions of Ethiopia: an interrupted time series analysis of health information system data from 2019–2020. PLOS Glob Public Health 2022;2(9):e0000843. [doi: 10.1371/journal.pgph.0000843]
75. Thakur G, Arora A, Sikka P, Jain V. Impact of covid 19 pandemic on severe maternal outcomes -an observational study from a referral institute of India. Clin Epidemiol Glob Health 2022;17:101121. [doi: 10.1016/j.cegh.2022.101121] [Medline: 35957952]

76. Vardell E. Global Health Observatory data repository. Med Ref Serv Q 2020;39(1):67-74. [doi: 10.1080/02763869.2019.1693231] [Medline: 32069199]
77. Hussein Sayed Abdel Gaowad A, Hamdio S, Fathi Nabaweya Saleh A. Effect of Covid19 pandemic on pregnant women utilization of antenatal care services. Egyptian Journal of Health Care 2022 Dec 1;13(4):668-681. [doi: 10.21608/ejhc.2022.265053]
78. Bonet M, Babinska M, Buekens P, et al. Maternal and perinatal health research during emerging and ongoing epidemic threats: a landscape analysis and expert consultation. BMJ Glob Health 2024 Mar 7;9(3):e014393. [doi: 10.1136/bmjgh-2023-014393] [Medline: 38453249]

Abbreviations

IV: intravenous
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PROSPERO: International Prospective Register of Systematic Reviews
RCT: randomized controlled trial
RR: rate ratio
UI: uncertainty interval

Edited by E Meinert, T Leung; submitted 21.02.24; peer-reviewed by Anonymous, IG Sastra Winata, S Kumareswaran; revised version received 11.07.25; accepted 31.07.25; published 06.10.25.

Please cite as:
Muthuka JK, Mbari-Fondo DK, Wambura FM, Oluoch K, Nzioki JM, Nyamai EM, Nabaweesi R
Effects of Interventions for the Prevention and Management of Maternal Anemia in the Advent of the COVID-19 Pandemic: Systematic Review and Meta-Analysis
JMIRx Med 2025;6:e57626
URL: https://xmed.jmir.org/2025/1/e57626
doi:10.2196/57626

Perception and Impact of White Spot Lesions in Young People Undergoing Orthodontic Treatment and Their Guardians: Protocol for a Mixed Methods Study

Amaar Obaid Hassan[1], BSc, BDS, MPH; Janine Doughty[2], BDS, MDPH, DDPH RCS Eng, PG Cert Clin Res, PGCAP, PhD; Jayne Harrison[3], BDS, MDentSci, PhD

[1]Department of Orthodontics, School of Dentistry, Liverpool University, Pembroke Place, Liverpool, United Kingdom

[2]School of Dentistry, University of Liverpool, Liverpool, United Kingdom

[3]Orthodontic Department, Liverpool University Dental Hospital, Liverpool, United Kingdom

Corresponding Author:
Amaar Obaid Hassan, BSc, BDS, MPH
Department of Orthodontics, School of Dentistry, Liverpool University, Pembroke Place, Liverpool, United Kingdom

Related Articles:

Companion article: http://preprints.jmir.org/preprint/60213

Companion article: https://med.jmirx.org/2025/1/e80143

Companion article: https://med.jmirx.org/2025/1/e80140

Companion article: https://med.jmirx.org/2025/1/e80139

Abstract

Background: White spot lesions (WSLs) are white marks that can form on teeth during orthodontic treatment with fixed appliances and become apparent once they are removed. About half of people who have fixed appliance treatment get WSLs. They are usually caused by poor toothbrushing around the brace. Although there have been studies that have investigated the prevention and treatment of WSL, there remain uncertainties about what young people and their parents or guardians know or feel about them. A Cochrane review concluded that patient-reported outcomes have been overlooked in WSL prevention studies.

Objective: The aim of this study is to explore young people's and their parents'/guardians' perceptions, attitudes, and feelings toward WSLs using a mixed methods study.

Methods: This is a mixed methods study. Part 1 is a cross-sectional survey using a web-based survey questionnaire and images of pretreatment malocclusions and postorthodontic WSLs of varying severity (mild, moderate, severe). Part 2 will involve one-to-one, semistructured interviews, using open-ended questions with young people and their parents/guardians. Participants will be recruited from patients aged 11 - 15 years before, during, or after undergoing orthodontic treatment at Liverpool University Dental Hospital and their parents/guardians. Part 1 (quantitative) will use a Likert scale with the option of free text comments. Data will be analyzed using descriptive statistics. Agreement between participants will be analyzed using the κ statistic. Part 2 (qualitative) will be analyzed using a modified framework analysis approach; the outcomes will be presented as themes. Transcripts from the qualitative interview will be analyzed using inductive thematic analysis. Once the qualitative and quantitative data have been analyzed, we will combine the two datasets and compare them for convergence or divergence. We will aim for a sample size of at least 100 participant and parent/guardian pairs for Part 1 and 30 interviewees for Part 2. Ethical approval was granted in November 2024. The Sponsor Permission to Proceed notification was received in January 2025.

Results: Funding for the study was secured in May 2024. Recruitment started on February 2, 2025. As of August 31st, 2025, seventy five participant pairs have been recruited.

Conclusions: The study will increase understanding of the impact WSLs have on oral health–related quality of life and the decision-making of young people and their parents/guardians.

(JMIRx Med 2025;6:e60213) doi:10.2196/60213

KEYWORDS

orthodontics; white spot lesions; fixed appliances; dentistry

Introduction

Background

White spot lesions (WSLs) are enamel defects that commonly appear as opaque, white, matte, chalky, or brown spots on teeth and can form around fixed orthodontic appliances. Approximately half of patients undergoing orthodontic treatment with fixed appliances will experience WSLs [1]. WSLs are caused by the combination of poor toothbrushing around the brackets of fixed appliances and frequent sugar/acidic attacks. A susceptible tooth surface exposed to bacterial plaque accumulation and fermentable carbohydrates over a sufficient period of time will undergo demineralization and potentially develop WSLs [2].

WSLs can reduce the quality and amount of enamel, leaving teeth vulnerable to damage by dental caries, thereby diminishing their lifetime prognosis [3]. Anterior surfaces on maxillary teeth are frequently affected (36%), thus WSLs may be highly visible when smiling or speaking [4].

Esthetic defects caused by WSLs may expose young people to oral health–related stigma and discrimination (eg, bullying or teasing) and impact self-esteem [5]. Low self-esteem is associated with lower grades at school, depression, and impaired social interaction with others [5]. Visible differences in dentofacial features such as tooth shape or color have been implicated as a driver for self-harm in teenagers [6]. Therefore, this study is not only important for preventing WSLs, but also for understanding the impact of WSLs on the oral health–related quality of life of young people.

There is evidence to suggest that fluoride can prevent WSLs by enhancing remineralization. Fluoride can be applied using various vehicles (eg, toothpaste, mouthwash, fluoride varnish, and casein phosphopeptides-amorphous calcium phosphate). High-strength fluoride toothpaste (5000 ppm) may be advantageous for preventing WSLs when compared with standard concentrations of fluoride toothpaste; however, these prescription-only toothpastes are only available from the age of 16 and many patients who have orthodontic treatment with fixed appliances are often below this age [7]. Once an orthodontic WSL is confirmed, then it may not be advantageous to expose it to high-strength fluoride as it can create a barrier to calcium and phosphate ions, meaning the lesion stains or persists [8].

Nonetheless, the most crucial aspect of WSL management is motivating patients to adhere to effective oral hygiene measures and noncariogenic dietary advice [9]. Clinicians believe the responsibility for preventing WSLs lies with the patient and that their postorthodontic outcomes are determined by their willingness to engage with the oral hygiene advice discussed at the commencement of treatment [10]. Visual aids have been found to be helpful to demonstrate the risks of WSLs and to motivate young people to maintain their oral hygiene [11]. A further challenge is presented by parents who are reluctant to assume responsibility for their child's oral hygiene practices [12].

Patient-reported outcomes have been overlooked in all studies exploring the efficacy of different interventions to prevent WSLs in patients undergoing fixed orthodontic treatment [7]. Patient perceptions of WSLs can provide insights into motivators and barriers to maintaining good oral health during orthodontic treatment. Additionally, WSLs may also have cost consequences for patient and National Health Service dental services, for example, the costs of professionally applied fluoride and cosmetic restorations and their long-term maintenance.

WSLs pose an important risk for patients when considering whether to opt for orthodontic treatment [13]. At present, clinicians may be negotiating these conversations without a full evidence-based understanding of patient perceptions toward WSLs.

What remains unknown, and is the focus of our proposed study, are patients' and parents' perceptions of WSLs including the impact of WSLs on the acceptability of orthodontic outcomes [7]. Therefore, this study offers the opportunity for researchers to identify ways to communicate WSL risk to patients and their parents and to understand how best to motivate good oral hygiene and dietary practices during orthodontic treatment with fixed appliances.

Reporting

Due to the absence of a well-recognized mixed methods reporting tool, we have chosen to use the Standards for Reporting Qualitative Research [14] Checklist for Reporting of Survey Studies guidelines [15] to assist with transparent and consistent reporting of each aspect of the study [16].

Aim

The aim of this study is to explore young people's and their parents/guardians' perceptions of and attitudes toward WSLs.

Objectives

The objectives of this study are to (1) create and use a questionnaire that assesses the perceptions and impact of WSLs on young people undergoing treatment with a fixed orthodontic appliance and their parents/guardians (see Multimedia Appendix 1) and (2) explore the impact and perceptions of WSL formation on young people undergoing fixed brace treatment and their parents/guardians using one-to-one interviews and visual images of malocclusions and WSLs of varying severity (see Multimedia Appendix 2).

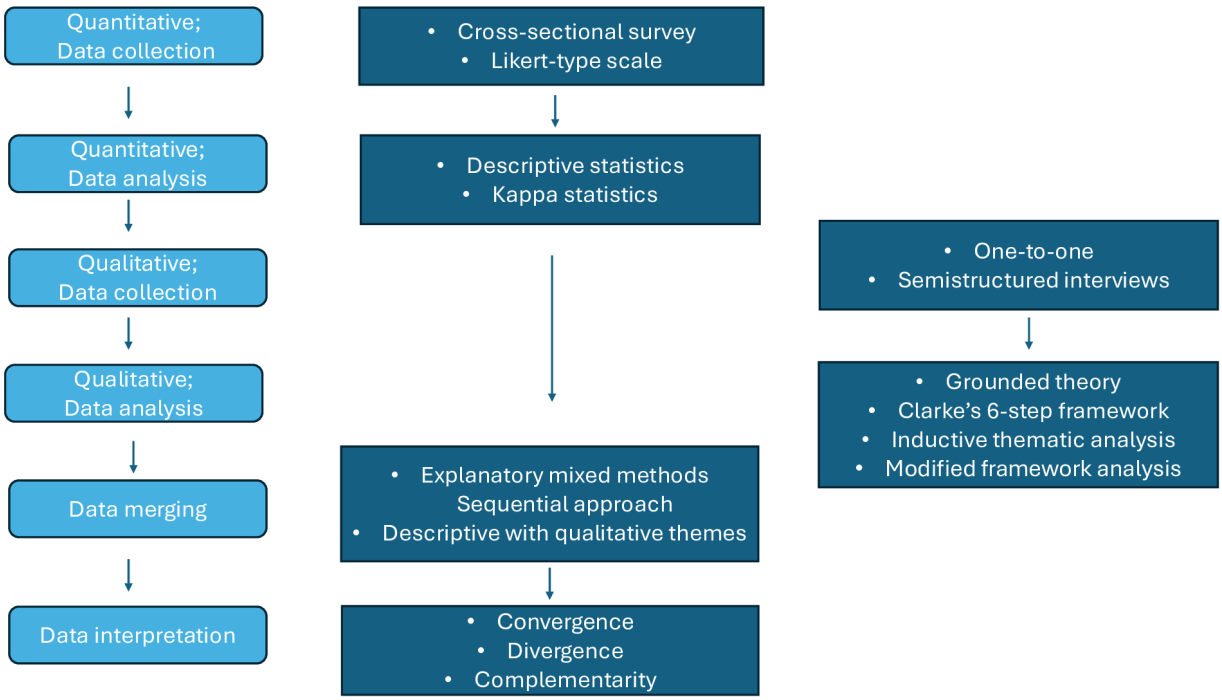
Methods

Summary

A mixed methods study design was chosen for this study to provide enriched data by augmenting the quantitative findings from a questionnaire (Part 1) with data from qualitative interviews (Part 2). The questionnaire was developed in collaboration with a patient and public involvement (PPI) group.

Following completion of the quantitative research (Part 1), meetings of the research team and the PPI group will be organized to review the data and develop the interview schedule for the qualitative phase (Part 2). The interview schedule will be based on the findings of Part 1 of the research. Thus, an explanatory sequential mixed methods approach will be used, whereby the qualitative data will expand on the understanding gained from the questionnaire [17]. The diagram below (Figure 1) illustrates the different parts of the study and at what point the mixing of the data will occur. Following completion of the data collection and analysis of both datasets, the results will be merged. The quantitative and qualitative data will then be compared for convergence and divergence.

Figure 1. Flowchart of the mixed methods design used in this study.



Part 1: Questionnaire

Quantitative data from a questionnaire will be collected and analyzed (Multimedia Appendix 1). The questionnaire was developed in collaboration with a PPI group, GenerationR, at Alder Hey Hospital, Liverpool, United Kingdom. The PPI group gave us insight into: (1) the features required to create an accessible questionnaire design, (2) the decision-making processes involved in assessing orthodontic risk and benefit, (3) the personal responsibility attributed to maintain oral hygiene, and (4) the value and importance of pretreatment oral hygiene instruction using accessible methods.

The questionnaire is in 2 sections and includes 5 questions about participant information and 15 questions about participants' perceptions of WSLs. A 5-point Likert-type scale will be used to assess participants' perceptions of WSLs. Although there are limited studies assessing young people's/parents' perception of WSLs in orthodontics using a Likert-type scale, other published studies in orthodontics have used similar approaches and sample sizes to those proposed in our protocol [18]. Close-ended questions have been used to elicit higher response from participants (Multimedia Appendix 1) [19]. Participants will also be asked to rate a selection of images of WSLs of varying levels of severity before and after treatment.

Part 2: Interviews

Qualitative data from one-to-one interviews will be collected to explore the findings from the questionnaire in more detail (Multimedia Appendix 2).

The study will use deductive (from the questionnaire responses) and inductive thematic analysis in a modified framework analysis approach that incorporates aspects of grounded theory [20]. The framework method will be used to gather data from the interviews for themes. It will follow Braun and Clarke's 6-step framework [21]:

1. Familiarization of the data.
2. Create codes.
3. Generate themes.
4. Review themes.
5. Define themes.
6. Write up the results.

Main themes and subthemes will be developed iteratively alongside further data collection, with a search for confirming and disconfirming cases, until data saturation is reached [22].

Research Team

AOH is the lead for the research. He is currently an orthodontic specialty trainee in the United Kingdom and is undertaking this project as part of a PhD degree. He has previous experience completing qualitative research and has an MRes degree.

JH is a professor and consultant in orthodontics in Liverpool, United Kingdom, and has a PhD degree. She is the Chief Investigator and primary supervisor for the project. She is currently involved in a randomized controlled trial looking at preventing WSLs (FL_4OWS or Fluorides for Orthodontic White Spots), for which she has a research grant.

JD has experience in PPI and mixed methods research [23]. She has a PhD degree and has highlighted the importance of this study with respect to oral health–related stigma and shame and is cosupervising the project.

Amy Rawsthorne, who is not included as an author of this paper, is a research nurse who has previous experience in completing qualitative research. She will be assisting with recruitment, obtaining consent, completion of the questionnaires, and one-to-one interviews.

As AOH is the lead for this study and will be recruiting participants and interviewing them, none of the participants in this study will be treated by him as it may influence their responses.

The University of Liverpool has a strong history of caries research in relation to orthodontic treatment including in vitro, in situ, and case control studies, as well as randomized controlled trials [24-26].

Context

The study will be undertaken in the orthodontic department at Liverpool University Dental Hospital, United Kingdom.

Sampling Strategy

Inclusion Criteria

Individuals were recruited if they met the following criteria:

- Young people aged 11 - 15 years inclusive who are considering or undergoing fixed orthodontic treatment at Liverpool University Dental Hospital, United Kingdom.
- Parent/guardian or person who has parental responsibility for a young person considering or undergoing fixed orthodontic treatment at Liverpool University Dental Hospital, United Kingdom.

Exclusion Criteria

Exclusion criteria were young people with:

- Learning difficulties that preclude them from answering the questionnaire or making an active contribution to interviews
- Craniofacial or other syndromes
- Nonorthodontic WSLs

For Part 1, the survey respondents will be recruited by convenience sampling methods from patients attending the orthodontic clinic at Liverpool University Dental Hospital, United Kingdom.

For Part 2, purposive sampling will be used to ensure the qualitative phase of the research recruits a sample with heterogeneous characteristics. Sampling will be based on age, gender, ethnicity, stage of treatment, and condition of first molar teeth based on previous clinical records (Table 1).

Table . Interview purposive sampling framework.[a]

		Gender	
		Male	Female
Functional appliance		Minimum 1	Minimum 1
Age (years)			
	11 - 13	Minimum 1	Minimum 1
	14 - 15	Minimum 1	Minimum 1
Ethnicity			
	Not Caucasian	Minimum 2	Minimum 2
Stage of treatment			
	Early	Minimum 1	Minimum 1
	Midtreatment	Minimum 1	Minimum 1
	Treatment completed	Minimum 1	Minimum 1
Condition of one or more first molars			
	Sound	Minimum 1	Minimum 1
	Restored/carious	Minimum 1	Minimum 1
	Extracted	Minimum 1	Minimum 1

[a]A minimum of 12 interview participants are needed.

Justification for This Framework

The sampling framework has been selected to represent a diverse sample and to limit bias. Differences condition of first molars are also considered because previous research within the department has shown that participants who experience WSLs have first molars in worse condition [27]; therefore, it would

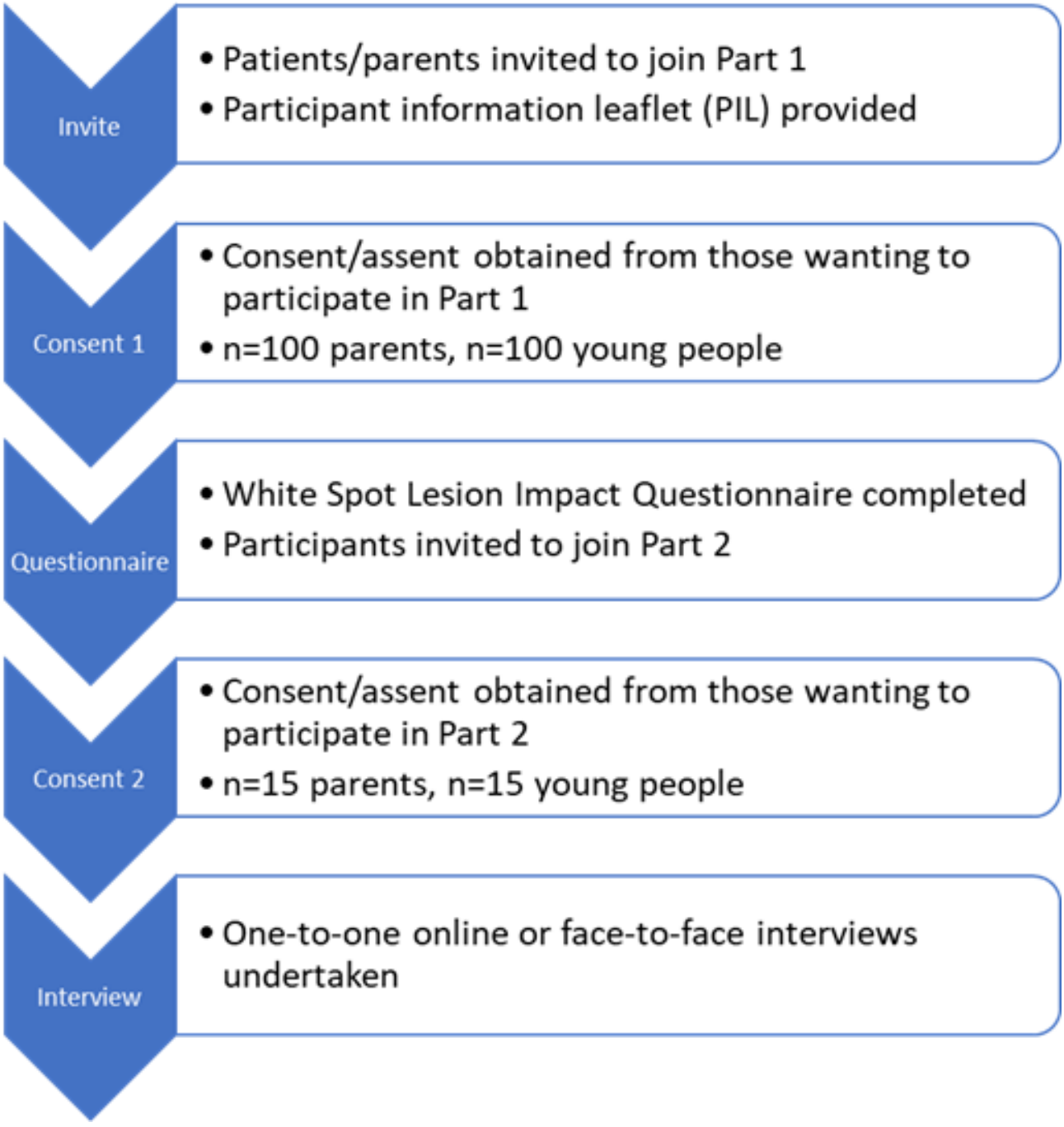
be useful to also collect participant data from this demographic as they are a higher risk group.

Data about the condition of first molars will be assessed using patient records and notes; participants will not be assessed by looking in their mouth or through taking radiographs specifically for this project.

Invitation

Eligible participants will be invited to take part in the research by a member of the research team (AOH). Eligible young people and their parents/guardians will be invited to complete the questionnaire on the same day as their routine orthodontic appointment. They will have free choice over whether they wish to take part and will be able to read the participant information leaflet during their appointment or after. Should they wish to take part in the study, they will be reimbursed for their time with a £10 (US $13.57) electronic voucher for the questionnaire and £25 (US $33.91) for the interviews. The flow of participants through the study is illustrated in Figure 2.

Figure 2. Participant journey during the study.



Sample Size

For Part 1, a pragmatic estimation of a sample size of 200 survey respondents (100 patients and 100 parents) had been initially agreed upon to address the aims and objectives of this part of the study. For a representative sample of patients aged 11 - 15 receiving routine orthodontic treatment at the department, with a population of approximately 550 patients, 95% confidence

level, and error margin of 5%, 226 respondents will be needed. Once data collection is near completion, the authors will ask a statistician to determine whether 226 respondents are needed.

For Part 2, for the qualitative research, 30 interviews with 15 young people and 15 parents/guardians will be undertaken. This sample size may be adjusted if data saturation is achieved earlier or not achieved and is comparable with the sample sizes of similar orthodontic studies [19,20].

Ethical Considerations

Ethical approval was granted, via the Integrated Research Approval System (333499), on November 8, 2024. Sponsor Permission to Proceed notification was received from University of Liverpool (UoL001871) on January 30, 2025. Recruitment started on February 2, 2025. Participants will have free choice over whether they wish to take part and will be able to read the participant information leaflet during their appointment or after. Patients willing to take part in the study will be provided with a password-protected tablet displaying an information leaflet about the questionnaire. The information can also be accessed using their own device via a QR code. For those willing to participate, written assent will be obtained from the participants under the age of 16 years, and written consent will be obtained from their parent/guardian. The main ethical issues are maintaining confidentiality and asking questions that might be of a sensitive nature to young people and their parents/guardians. These will be addressed by strict compliance with institutional protocols about anonymization and storage of data. The young people will be interviewed separately from their parents/guardians to avoid corroboration. A chaperone will be present during the interviews with the young people, if requested. Participants will be able to stop or pause the interview at any stage. Should individuals wish to take part in the study, they will be reimbursed for their time with a £10 (US $13.57) electronic voucher for the questionnaire and £25 (US $33.91) for the interviews.

Data Collection Methods, Instruments, and Technologies

Part 1: Quantitative Survey

Data collection for the survey started on February 2, 2025, and will continue until the sample size is met. Participant demographic data and the responses to the questions will be collected.

The participants will complete the questionnaire on a password-protected tablet in a private room near the orthodontic clinic at Liverpool University Dental Hospital. The questionnaire will be accessed using JISC Online Surveys through the University of Liverpool. Data will be transferred from the questionnaire using JISC and exported to an Excel 2018 (Microsoft Corp) spreadsheet for analysis.

The research team will review the data after 25% of the sample has completed the questionnaire to check if participants who have been recruited are representative of the demographics of the clinic. If any groups of people are underrepresented at this point, then this will be identified, and efforts will be made to recruit people from the underrepresented groups.

Part 2: Qualitative Survey

Data collection will start after Part 1 is complete and the data have been analyzed. It will continue until data saturation is achieved. Audio data will be transcribed and saved for analysis using NVivo software (version 15; Lumivero).

The young people will be interviewed separately from their parents/guardians to ensure discussions are not influenced by each other. If required, a chaperone will be present for the young people. A semistructured one-to-one interview process will enable the researcher to explore relevant topics raised by the participants.

The qualitative interviews will be conducted remotely via videoconference or in a private room in the orthodontic clinic at Liverpool University Dental Hospital, United Kingdom, using Teams. Audio recordings will be autotranscribed by Teams and checked and edited by a university-approved transcription service or AOH. Once transcribed, the audio recording will be deleted.

An interview guide/schedule has been devised to provide flexible direction and consistency and to probe all key topics sufficiently. The interview schedule will be adapted to take into account the findings of Part 1 of the study (Multimedia Appendix 2) and following a further PPI meeting to develop the interview schedule further. Discussions will last for approximately 45 - 60 minutes. Interviews will continue until there are no recurring themes or patterns emerging from the data and thematic saturation has been achieved.

Field Notes

Written notes will be taken throughout the research to document behaviors and communications that were not present in the audio. Field notes will also be useful to allow researchers to write reflections that will help in analyzing the data.

Units of Study

The relevant inclusion and exclusion criteria will be applied as mentioned previously.

The target sample size will be 226 (113 patient and parent/guardian pairs) for Part 1 and 30 (15 patient and parent/guardian pairs) for Part 2 or until data saturation is reached.

Data Processing

For Part 1, data from the questionnaire will be collected directly onto a password-protected tablet. Data will be transferred to and stored on a password-protected University of Liverpool or National Health Service computer during collection and for analysis. Data will be anonymized and analyzed using participant ID numbers. Microsoft Excel will be used for the descriptive data analysis.

For Part 2, interviews will be recorded directly onto a password-protected tablet. Transcription will be undertaken automatically via Teams and checked by a University of Liverpool provider or AOH. If needed, transcripts will be sent to participants for any clarification. NVivo software will be used to analyze the interview data.

Data will be retained in the University of Liverpool secure server for 10 years and deleted after. There will be no sharing of the data outside the immediate research team.

Data Analysis

For Part 1, descriptive statistics will be used to analyze the data from the questionnaire. The κ statistic will be used to measure the agreement of responses between participants and their parents/guardians.

For Part 2, during the qualitative research analysis, data coding and identification of themes of transcripts will be undertaken by AOH using NVivo 12. The transcripts and codes/themes generated will be sent to a second or third researcher to confirm reliability (JH, JD, or AR). If needed, transcripts will be sent to participants for any clarification.

The study will use deductive (from the questionnaire responses) and inductive thematic analysis in a modified framework analysis approach that incorporates aspects of grounded theory [20]. The framework method will be used to gather data from the interviews for themes. It will follow Braun and Clarke's 6-step framework [21]. Main themes and subthemes will be developed iteratively alongside further data collection, with a search for confirming and disconfirming responses, until data saturation is reached [15]. A coding tree will illustrate and organize the qualitative data to identify different themes.

Once both individual sets of data have been analyzed, they will be combined and compared for convergence or divergence, similar to another study published previously [28].

Techniques to Enhance Trustworthiness

Data and methodological triangulation will be achieved by collecting data from a questionnaire and interviews to study the same phenomenon and allow researchers to cross-validate the findings. AOH will collect the data. JH and JD will be involved with data analysis and interpretation to ensure investigator triangulation. AOH will also maintain field notes of findings not recorded by the audio to allow for reflection.

To ensure member checking, the GenerationR PPI group will be involved in interpreting the Part 1 results and developing the interview schedule. Participants who have agreed to join this group will be contacted following data analysis to check whether they are happy with the interpretation and conclusions of the study and if they have any further comments to add. The research team will also present findings of the study to the PPI group to ensure their input into the interpretation of the findings.

AOH and AR will maintain an audit trail of the study documentation in line with Trust policy.

The research team will reflect on their own biases and assumptions throughout the research process and try to remain neutral during all stages of data collection, analysis, and interpretation. AOH will avoid recruitment of participants who are under his direct clinical care.

Peer review will be sought through presenting the results at local and national meetings prior to publication.

Duration

A pilot of the study invited 10 young people and 10 parents/guardians to read the questionnaire and provide feedback on it. The authors were also able to monitor the length of time it took to invite and recruit people to the study. Following the pilot, an amendment to the time needed for recruitment was submitted to the sponsor and National Health Service ethics committee to allow for enough time for data collection for both parts of the study. A Gantt chart was created to provide further details on the duration of the study (Table 2).

Table . Gantt chart illustrating timelines for this white spot lesion research.

	Jan-25	Feb-25	Mar-25	Apr-25	May-25	Jun-25	Jul-25	Aug-25	Sep-25	Oct-25	Nov-25	Dec-25	Jan-26	Feb-26	Mar-26
Recruitment for quantitative research	✓	✓	✓	✓	✓	✓	✓	✓							
Data collection for quantitative research	✓	✓	✓	✓	✓	✓	✓	✓							
Data analysis of quantitative research								✓							
Recruitment for qualitative interviews									✓	✓	✓	✓	✓	✓	✓
Transcription of interviews									✓	✓	✓	✓	✓	✓	✓
Data analysis of interviews												✓	✓	✓	✓

Results

Progress to March 13, 2025

Funding for the study was secured through a joint Faculty of Dental Surgery and Royal College of Surgeons of England/British Orthodontic Society pump-priming grant in May 2024. Ethical approval was granted via the Integrated Research Approval System (333,499) on November 8, 2024. The Sponsor Permission to Proceed notification was received from the University of Liverpool (UoL001871) on January 30, 2025. Recruitment started on February 2, 2025. As of August 31st, 2025, seventy five participant pairs have been recruited. Results are expected to be published in a peer-reviewed journal in the summer of 2026.

Part 1

Demographic data of the sample and descriptive data for the responses from the questionnaire together with associations between participants and the young people and parent/guardian pairs will be presented.

Part 2

Themes and subthemes will be explored; interpretations and inferences will be presented, including links to empirical data from the interview transcripts.

Discussion

Overview

In the final paper, a summary of the findings will be presented. The results will be compared to the existing literature. The strengths and limitations of the study and data will be explored. The implications for clinical practice and further research will be discussed.

Strengths

The aims of orthodontic treatment are to improve the occlusion and appearance of teeth, which benefits function, confidence, self-esteem, and quality of life [23,29]. Detecting WSLs after removing fixed orthodontic appliances may detract from the benefits of orthodontic treatment. WSLs are likely to have negative associations for people with anterior tooth discoloration, and they may experience unfavorable judgments of personality traits and characteristics, which may affect friendships, relationships, and career prospects [30].

WSLs also have implications for the consent process for orthodontic treatment. Although young people can understand risks, they can have false perceptions about them if they are not fully understood [31]. Poor communication can lead to issues with consent and other negative outcomes like complaints and litigation [31].

This mixed methods study has the potential to inform clinicians' communication about WSLs with young people and their parents/guardians. Furthermore, the study may help researchers improve their understanding of what methods can help to inform young people of the potential consequences of WSLs and support WSL prevention. Even with effective oral hygiene instruction, around half of young people do not follow the clinician's advice to improve their oral hygiene [12]. The COM-B model is presented as a tool to diagnose which of capability, opportunity, or motivation need to change for a new behavior to take place [32]. Although interventions designed to improve oral hygiene during orthodontic treatment (including using smartphones and a toothbrushing app, visual aids, motivational interviewing, oral health reinforcements) have been investigated, only the use of mobile phones has had limited evidence for improving oral health during orthodontic treatment [33]. To our knowledge, studies have not been undertaken to explore barriers to oral hygiene or behavioral interventions to reduce WSL formation during orthodontic treatment in young people.

Limitations

The study is only recruiting participants from one hospital in Liverpool, United Kingdom, and there may be a difference if the young people and parents/guardians were to be recruited in primary care or from another area of the United Kingdom and/or another country. Patients attending hospital orthodontic departments tend to have more complex treatment needs, which might influence their perceived posttreatment satisfaction.

With all cross-sectional studies, there are limitations to questionnaires, as they collect data at a set time point and are therefore unable to establish cause and effect. The participants who are likely to respond to the questionnaire are more likely to be young people and their parents/guardians who are interested in the project and may be more motivated to prevent orthodontic WSLs than those who do not volunteer to take part. Although it will be difficult to ensure all relevant groups in the population are included, the study will recruit a heterogeneous sample of participants with regard to age, gender, condition of first molars, and type of orthodontic treatment they are receiving. The study will identify participants from different cells of a sampling framework (Table 1). Social desirability bias has been limited by asking participants not to discuss answers with parents/guardians as this may influence their answers. The participants are able to complete the questionnaire in a private room without a researcher being present. The study will not recruit any participants who are under the clinical care of the research team involved in recruiting. PPI will be used throughout all stages of the research to ensure questions are relevant to the participants and not misleading.

As a clinician/dentist, AOH will undertake the qualitative research; this increases the risk that the researcher could make assumptions about what the participants think or feel based on their professional experience. Parents/guardians may not want to feel blamed by health care professionals and may want to avoid being responsible for the young person's poor oral hygiene practices. Participants may forget to recall information during the interview, and all participants could answer questions differently based on what they think the researcher would like to hear as the "correct" answer rather than discussing their own honest experience, especially if they are aware that the researcher is a clinician. As discussed, the study will confirm the reliability of the answers to the questionnaire and qualitative interviews. To aid transparency, if there is a disagreement in the interpretation of the qualitative data, the themes/codes will be sent to another researcher for secondary/tertiary analysis. Where possible, AOH will avoid recruiting patients directly under his care, as it may affect their responses and/or his questions due to prior knowledge of each other. The study will also promote reflexivity through a postinterview debrief with participants and members of the research team. This will help to ensure that the study findings agree with the participant views rather than any subjectivity or researcher bias. Participants also can review study findings to ensure that they agree with the results. The authors have also attempted to address self-reporting bias by publishing the study protocol, the questionnaire/interview schedule, and the data so that readers are able to make an informed decision about the potential sources of bias.

Conclusions

This is a mixed methods study that aims to investigate the impact of WSLs on young people who are considering or undergoing orthodontic treatment with fixed appliances and their parent/guardian, as well as to explore their perceptions, attitudes, and feelings toward WSLs.

Acknowledgments

Funding for the study was secured through a joint Faculty of Dental Surgery, Royal College of Surgeons of England/British Orthodontic Society pump-priming grant of £7693.70 (US $10,430) in May 2024. The funding is to provide vouchers for

participants and members of the PPI group as well as for printing and external transcription costs and an iPad to assist with data collection. The external funding does not include the salary of staff or researchers.

Data Availability

Any data generated/analyzed not presented in the paper are available on request.

Authors' Contributions

AOH, JH, JD were all involved in writing up the article. AOH will be involved in recruitment and data collection. AOH, JH, and JD will all be involved in data analysis.

Conflicts of Interest

AOH is an orthodontic specialty trainee, so the interviews will be influenced by his prior knowledge, attitudes, and position. He is not undertaking the treatment of any of the participants, so the impact of patients knowing him will be avoided. JH has been awarded funding from the British Orthodontic Society Foundation for a randomized controlled trial looking at two different clinical interventions for preventing WSLs in young people with fixed orthodontic appliances. She is not receiving any reimbursements, fees, funding, or salary from the funding organization and will not in any way personally gain or lose financially from the publication of the manuscript.

Multimedia Appendix 1
Questionnaire.
[DOCX File, 732 KB - xmed_v6i1e60213_app1.docx]

Multimedia Appendix 2
Interview schedule.
[DOCX File, 18 KB - xmed_v6i1e60213_app2.docx]

References

1. Gorelick L, Geiger AM, Gwinnett AJ. Incidence of white spot formation after bonding and banding. Am J Orthod 1982 Feb;81(2):93-98. [doi: 10.1016/0002-9416(82)90032-x] [Medline: 6758594]
2. Srivastava K, Tikku T, Khanna R, Sachan K. Risk factors and management of white spot lesions in orthodontics. J Orthod Sci 2013 Apr;2(2):43-49. [doi: 10.4103/2278-0203.115081] [Medline: 24987641]
3. Braga MM, Mendes FM, Ekstrand KR. Detection activity assessment and diagnosis of dental caries lesions. Dent Clin North Am 2010 Jul;54(3):479-493. [doi: 10.1016/j.cden.2010.03.006] [Medline: 20630191]
4. Chapman JA, Roberts WE, Eckert GJ, Kula KS, González-Cabezas C. Risk factors for incidence and severity of white spot lesions during treatment with fixed orthodontic appliances. Am J Orthod Dentofacial Orthop 2010 Aug;138(2):188-194. [doi: 10.1016/j.ajodo.2008.10.019] [Medline: 20691360]
5. Seehra J, Fleming PS, Newton T, DiBiase AT. Bullying in orthodontic patients and its relationship to malocclusion, self-esteem and oral health-related quality of life. J Orthod 2011 Dec;38(4):247-256. [doi: 10.1179/14653121141641] [Medline: 22156180]
6. Al-Bitar ZB, Sonbol HN, Al-Omari IK, et al. Self-harm, dentofacial features, and bullying. Am J Orthod Dentofacial Orthop 2022 Jul;162(1):80-92. [doi: 10.1016/j.ajodo.2021.02.025] [Medline: 35346538]
7. Benson PE, Parkin N, Dyer F, Millett DT, Germain P. Fluorides for preventing early tooth decay (demineralised lesions) during fixed brace treatment. Cochrane Database Syst Rev 2019 Nov 17;2019(11):CD003809. [doi: 10.1002/14651858.CD003809.pub4] [Medline: 31742669]
8. Ogaard B, Rølla G, Arends J, ten Cate JM. Orthodontic appliances and enamel demineralization. Part 2. Prevention and treatment of lesions. Am J Orthod Dentofacial Orthop 1988 Aug;94(2):123-128. [doi: 10.1016/0889-5406(88)90360-5] [Medline: 3165239]
9. Lopatiene K, Borisovaite M, Lapenaite E. Prevention and treatment of white spot lesions during and after treatment with fixed orthodontic appliances: a systematic literature review. J Oral Maxillofac Res 2016;7(2):e1. [doi: 10.5037/jomr.2016.7201] [Medline: 27489605]
10. Maxfield BJ, Hamdan AM, Tüfekçi E, Shroff B, Best AM, Lindauer SJ. Development of white spot lesions during orthodontic treatment: perceptions of patients, parents, orthodontists, and general dentists. Am J Orthod Dentofacial Orthop 2012 Mar;141(3):337-344. [doi: 10.1016/j.ajodo.2011.08.024] [Medline: 22381494]
11. Sundararaj D, Venkatachalapathy S, Tandon A, Pereira A. Critical evaluation of incidence and prevalence of white spot lesions during fixed orthodontic appliance treatment: a meta-analysis. J Int Soc Prev Community Dent 2015;5(6):433-439. [doi: 10.4103/2231-0762.167719] [Medline: 26759794]

12. Mei L, Chieng J, Wong C, Benic G, Farella M. Factors affecting dental biofilm in patients wearing fixed orthodontic appliances. Prog Orthod 2017 Dec;18(1):4. [doi: 10.1186/s40510-016-0158-5] [Medline: 28133715]
13. Perry J, Popat H, Johnson I, Farnell D, Morgan MZ. Professional consensus on orthodontic risks: what orthodontists should tell their patients. Am J Orthod Dentofacial Orthop 2021 Jan;159(1):41-52. [doi: 10.1016/j.ajodo.2019.11.017] [Medline: 33221095]
14. O'Brien BC, Harris IB, Beckman TJ, Reed DA, Cook DA. Standards for Reporting Qualitative Research: a synthesis of recommendations. Acad Med 2014 Sep;89(9):1245-1251. [doi: 10.1097/ACM.0000000000000388] [Medline: 24979285]
15. Sharma A, Minh Duc NT, Luu Lam Thang T, et al. A consensus-based Checklist for Reporting of Survey Studies (CROSS). J Gen Intern Med 2021 Oct;36(10):3179-3187. [doi: 10.1007/s11606-021-06737-1] [Medline: 33886027]
16. Tong A, Sainsbury P, Craig J. Consolidated criteria for reporting qualitative research (COREQ): a 32-item checklist for interviews and focus groups. Int J Qual Health Care 2007 Dec;19(6):349-357. [doi: 10.1093/intqhc/mzm042] [Medline: 17872937]
17. Schoonenboom J, Johnson RB. How to construct a mixed methods research design. Kolner Z Soz Sozpsychol 2017;69(Suppl 2):107-131. [doi: 10.1007/s11577-017-0454-1] [Medline: 28989188]
18. Li Y, Liu J, Xu Y, Yin J, Li L. Oral health self-management ability and its influencing factors among adolescents with fixed orthodontics in China: a mixed methods study. Dis Markers 2022 Aug 27;2022(3657357):1-8. [doi: 10.1155/2022/3657357]
19. Griffith LE, Cook DJ, Guyatt GH, Charles CA. Comparison of open and closed questionnaire formats in obtaining demographic information from Canadian general internists. J Clin Epidemiol 1999 Oct;52(10):997-1005. [doi: 10.1016/s0895-4356(99)00106-7] [Medline: 10513763]
20. Ramanadhan S, Revette AC, Lee RM, Aveling EL. Pragmatic approaches to analyzing qualitative data for implementation science: an introduction. Implement Sci Commun 2021 Jun 29;2(1):70. [doi: 10.1186/s43058-021-00174-1] [Medline: 34187595]
21. Braun V, Clarke V. Using thematic analysis in psychology. Qual Res Psychol 2006 Jan;3(2):77-101. [doi: 10.1191/1478088706qp063oa]
22. Kiger ME, Varpio L. Thematic analysis of qualitative data: AMEE Guide No. 131. Med Teach 2020 Aug;42(8):846-854. [doi: 10.1080/0142159X.2020.1755030] [Medline: 32356468]
23. Doughty J, Macdonald ME, Muirhead V, Freeman R. Oral health-related stigma: describing and defining a ubiquitous phenomenon. Community Dent Oral Epidemiol 2023 Dec;51(6):1078-1083. [doi: 10.1111/cdoe.12893] [Medline: 37462247]
24. Benson PE, Pender N, Higham SM. An in situ caries model to study demineralisation during fixed orthodontics. Clin Orthod Res 1999 Aug;2(3):143-153. [doi: 10.1111/ocr.1999.2.3.143] [Medline: 10534989]
25. Doherty UB, Benson PE, Higham SM. Fluoride-releasing elastomeric ligatures assessed with the in situ caries model. Eur J Orthod 2002 Aug;24(4):371-378. [doi: 10.1093/ejo/24.4.371] [Medline: 12198867]
26. Garry AP, Flannigan NL, Cooper L, Komarov G, Burnside G, Higham SM. A randomised controlled trial to investigate the remineralising potential of Tooth Mousse in orthodontic patients. J Orthod 2017 Sep;44(3):147-156. [doi: 10.1080/14653125.2017.1341729] [Medline: 28681698]
27. Al Maaitah EF, Adeyemi AA, Higham SM, Pender N, Harrison JE. Factors affecting demineralization during orthodontic treatment: a post-hoc analysis of RCT recruits. Am J Orthod Dentofacial Orthop 2011 Feb;139(2):181-191. [doi: 10.1016/j.ajodo.2009.08.028] [Medline: 21300246]
28. Gaio DC, Bastos FI, Moysés SJ, et al. Assessing oral health of crack users in Brazil: perceptions and associated factors, findings from a mixed methods study. Glob Public Health 2021 Apr;16(4):502-516. [doi: 10.1080/17441692.2020.1809693] [Medline: 32912074]
29. Johal A, Alyaqoobi I, Patel R, Cox S. The impact of orthodontic treatment on quality of life and self-esteem in adult patients. Eur J Orthod 2015 Jun;37(3):233-237. [doi: 10.1093/ejo/cju047] [Medline: 25214505]
30. Kershaw S, Newton JT, Williams DM. The influence of tooth colour on the perceptions of personal characteristics among female dental patients: comparisons of unmodified, decayed and 'whitened' teeth. Br Dent J 2008 Mar 8;204(5):E9. [doi: 10.1038/bdj.2008.134] [Medline: 18297050]
31. Perry J, Johnson I, Popat H, Morgan MZ, Gill P. Adolescent perceptions of orthodontic treatment risks and risk information: a qualitative study. J Dent 2018 Jul;74:61-70. [doi: 10.1016/j.jdent.2018.04.011] [Medline: 29702151]
32. Michie S, van Stralen MM, West R. The behaviour change wheel: a new method for characterising and designing behaviour change interventions. Implement Sci 2011 Apr 23;6(42):42. [doi: 10.1186/1748-5908-6-42] [Medline: 21513547]
33. Farhadifard H, Soheilifar S, Farhadian M, Kokabi H, Bakhshaei A. Orthodontic patients' oral hygiene compliance by utilizing a smartphone application (Brush DJ): a randomized clinical trial. BDJ Open 2020 Nov 20;6(1):24. [doi: 10.1038/s41405-020-00050-5] [Medline: 33298841]

Abbreviations

FA: fixed appliance
PPI: patient and public involvement

WSL: white spot lesion

Edited by E Meinert, T Leung; submitted 04.05.24; peer-reviewed by A Jamilian, J Shaw; revised version received 19.03.25; accepted 03.07.25; published 12.09.25.

Please cite as:
Hassan AO, Doughty J, Harrison J
Perception and Impact of White Spot Lesions in Young People Undergoing Orthodontic Treatment and Their Guardians: Protocol for a Mixed Methods Study
JMIRx Med 2025;6:e60213
URL: https://xmed.jmir.org/2025/1/e60213
doi:10.2196/60213

Peer Review for “Applications of Indocyanine Green in Breast Cancer for Sentinel Lymph Node Mapping: Protocol for a Scoping Review”

Anonymous

Related Articles:

Companion article: https://www.medrxiv.org/content/10.1101/2024.07.30.24311256v1

Companion article: https://med.jmirx.org/2025/1/e68769

Companion article: https://med.jmirx.org/2025/1/e66213

(JMIRx Med 2025;6:e69705) doi:10.2196/69705

KEYWORDS

indocyanine green; ICG; sentinel lymph node; breast cancer; breast; fluorescence; axillary lymph node mapping; NIR; surgical planning; near-infrared

This is the peer-review report for “Applications of Indocyanine Green in Breast Cancer for Sentinel Lymph Node Mapping: Protocol for a Scoping Review.”

Round 1 Review

General Comments

This paper [1] summarized the application value and existing problems of indocyanine green (ICG) in sentinel lymph node (SLN) biopsy of early breast cancer, which has positive significance for improving the accuracy of clinical SLN detection. This study has certain clinical value.

Specific Comments

Major Comments

1. Due to the high hardware requirements for the clinical application of ICG, the number of relevant studies in the search is relatively small. It is hoped that the author can search the recent relevant literature to improve the credibility of this review.
2. It is hoped that the author will analyze and compare the advantages and disadvantages of ICG and traditional SLN biopsy methods, so as to guide clinicians to adopt appropriate methods for appropriate patients.

Conflicts of Interest

None declared.

Reference

1. Kurdi F, Kurdi Y, Reshetov IV. Applications of indocyanine green in breast cancer for sentinel lymph node mapping: protocol for a scoping review. JMIRx Med 2024;5:e66213. [doi: 10.2196/66213]

Abbreviations

ICG: indocyanine green
SLN: sentinel lymph node

Edited by S Tungjitviboonkun; submitted 05.12.24; this is a non–peer-reviewed article;accepted 05.12.24; published 06.01.25.

Please cite as:
Anonymous
Peer Review for "Applications of Indocyanine Green in Breast Cancer for Sentinel Lymph Node Mapping: Protocol for a Scoping Review"
JMIRx Med 2025;6:e69705
URL: https://xmed.jmir.org/2025/1/e69705
doi:10.2196/69705

Peer Review of "The Impact of Rural Alimentation on the Motivation and Retention of Indigenous Community Health Workers in India: Qualitative Study"

Sanjeev Kumar Thalari, MBA, MSc, PhD

Department of Management Studies & Research Center, CMR Institute of Technology, Bangalore, India

Related Articles:

Companion article: https://www.medrxiv.org/content/10.1101/2023.04.12.23288461v1

Companion article: https://med.jmirx.org/2025/1/e70059

Companion article: https://med.jmirx.org/2025/1/e48346

(JMIRx Med 2025;6:e70808) doi:10.2196/70808

KEYWORDS

rural alimentation; community health workers; motivation; retention; rural health; rural nutrition; workforce

This is the peer-review report for "The Impact of Rural Alimentation on the Motivation and Retention of Indigenous Community Health Workers in India: Qualitative Study."

Round 1 Review

General Comments

This paper [1] has given the impression that the researcher has done thorough homework before starting the research and it is evident in the paper. Case methodology and thematic analysis are a few of the approaches that depict the quality of the paper. Overall, as a reviewer, it is my opinion that the research paper is of quality.

Specific Comments

1. A few more factors like government initiatives should be included in studying the impact on the motivation and retention of community health workers.

Major Comments

1. I feel that the analysis also can include education as a parameter.

2. The thematic analysis is one of the strengths of this research and is appreciated.

Minor Comments

1. Common wording should be used in every section of the paper, like qualitative case research methodology and qualitative case research.

Conflicts of Interest

None declared.

Reference

1. Kerketta A, A N R. The impact of rural alimentation on the motivation and retention of Indigenous community health workers in India: qualitative study. JMIRx Med 2025;6:e48346. [doi: 10.2196/48346]

Edited by A Schwartz; submitted 02.01.25; this is a non–peer-reviewed article;accepted 02.01.25; published 23.01.25.

Please cite as:
Kumar Thalari S
Peer Review of "The Impact of Rural Alimentation on the Motivation and Retention of Indigenous Community Health Workers in India: Qualitative Study"
JMIRx Med 2025;6:e70808
URL: https://xmed.jmir.org/2025/1/e70808
doi:10.2196/70808

Peer Review of "Identifying Safeguards Disabled by Epstein-Barr Virus Infections in Genomes From Patients With Breast Cancer: Chromosomal Bioinformatics Analysis"

Anonymous

Related Articles:

Companion article: https://www.medrxiv.org/content/10.1101/2023.07.03.23292185v1

Companion article: https://med.jmirx.org/2025/1/e69307

Companion article: https://med.jmirx.org/2025/1/e50712

(JMIRx Med 2025;6:e70039) doi:10.2196/70039

KEYWORDS

breast; cancer; oncology; ovarian; virus; viral; Epstein-Barr; herpes; bioinformatics; chromosome; gene; genetic; genetics; chromosomal; DNA; genomic; BRCA; metastasis; biology

This is the peer-review report for "Identifying Safeguards Disabled by Epstein-Barr Virus Infections in Genomes From Patients With Breast Cancer: Chromosomal Bioinformatics Analysis."

Round 1 Review

Review Report With Major Revisions for the Paper

Title: "Herpesvirus infections eliminate safeguards against breast cancer and its metastasis: comparable to hereditary breast cancers"

Summary

The paper [1] hypothesizes that Epstein-Barr virus (EBV) infections promote breast cancer by disabling cancer safeguards. It is a bioinformatics analysis of public information from about 2100 breast cancers. The study finds that breast and ovarian cancer breakpoints cluster around EBV-associated cancer breakpoints, suggesting a significant role of EBV in promoting these cancers. The paper also identifies similarities in the molecular and cellular disruptions caused by EBV with those found in hereditary breast cancers.

Major Revisions Needed

Clarification of Hypotheses and Objectives

The hypothesis, while intriguing, needs clearer articulation. Specifically, the connection between EBV and breast cancer needs more explicit theoretical underpinning. Clarify the objectives and expected outcomes of the study at the outset.

Methodological Rigor and Data Sources

While the bioinformatics approach is robust, it would benefit from a more detailed description of the methods and algorithms used. Additionally, the selection criteria for the breast cancer data should be justified more thoroughly to avoid selection bias.

Statistical Analysis

The statistical methods used need more comprehensive detailing. For complex analyses, ensure the statistical assumptions and any transformations of data are clearly explained. Include more information on the statistical tests used for hypothesis testing and the justification for their use.

Comparative Analysis

The comparison between hereditary breast cancers and those potentially caused by EBV is insightful. However, a more detailed comparative analysis would strengthen the argument. This could include molecular or genetic profiling comparisons.

Discussion on Contradictory or Supporting Evidence

The discussion section should address not only the supporting evidence but also any contradictory findings in the literature. This balance is crucial for a nuanced understanding of the subject.

Implications and Future Research Directions

The implications of these findings are profound but need clearer articulation. Discuss the potential impact on breast cancer treatment and prevention strategies. Also, outline future research directions, particularly in clinical or experimental studies, to confirm these bioinformatics findings.

References

Please add more background information about breast cancer (please cite: 1. Cao Y, Efetov S, He M, et al. Updated clinical perspectives and challenges of chimeric antigen receptor-T cell therapy in colorectal cancer and invasive breast cancer. Arch Immunol Ther Exp (Warsz). Aug 11, 2023;71(1):19. [doi: 10.1007/s00005-023-00684-x] [Medline: 37566162]; and 2. Liu Y, Lu S, Sun Y, et al. Deciphering the role of QPCTL in glioma progression and cancer immunotherapy. Front Immunol.

Mar 29, 2023;14:1166377. [doi: 10.3389/fimmu.2023.1166377] [Medline: 37063864]).

Concluding Remarks

The paper presents a novel and potentially significant hypothesis linking EBV to breast cancer. However, it requires major revisions to enhance its methodological rigor, clarity, and comprehensiveness. Addressing these concerns will significantly strengthen the manuscript's impact and contribution to the field.

Conflicts of Interest

None declared.

Reference

1. Friedenson B. Identifying safeguards disabled by Epstein-Barr virus infections in genomes from patients with breast cancer: chromosomal bioinformatics analysis. JMIRx Med 2025;6:e50712. [doi: 10.2196/50712]

Abbreviations

EBV: Epstein-Barr virus

Edited by A Schwartz; submitted 13.12.24; this is a non–peer-reviewed article;accepted 13.12.24; published 29.01.25.

Please cite as:
Anonymous
Peer Review of "Identifying Safeguards Disabled by Epstein-Barr Virus Infections in Genomes From Patients With Breast Cancer: Chromosomal Bioinformatics Analysis"
JMIRx Med 2025;6:e70039
URL: https://xmed.jmir.org/2025/1/e70039
doi:10.2196/70039

Peer Review of "Identifying Safeguards Disabled by Epstein-Barr Virus Infections in Genomes From Patients With Breast Cancer: Chromosomal Bioinformatics Analysis"

Anonymous

Related Articles:

Companion article: https://www.medrxiv.org/content/10.1101/2023.07.03.23292185v1

Companion article: https://med.jmirx.org/2025/1/e69307

Companion article: https://med.jmirx.org/2025/1/e50712

(JMIRx Med 2025;6:e70041) doi:10.2196/70041

KEYWORDS

breast; cancer; oncology; ovarian; virus; viral; Epstein-Barr; herpes; bioinformatics; chromosome; gene; genetic; chromosomal; DNA; genomic; BRCA; metastasis; biology

This is the peer-review report for "Identifying Safeguards Disabled by Epstein-Barr Virus Infections in Genomes From Patients With Breast Cancer: Chromosomal Bioinformatics Analysis."

Round 1 Review

Dear Author,

After a thorough review of the paper titled "Herpesvirus infections eliminate safeguards against breast cancer and its metastasis: comparable to hereditary breast cancers" [1] by Bernard Friedenson, here is the negative feedback and evaluation, along with a recommendation for the inclusion of a specific article in the discussion section.

Negative Feedback and Evaluation

Clarity and Scope

The paper ambitiously attempts to link Epstein-Barr virus (EBV) infections to breast cancer development and metastasis. While the hypothesis is intriguing, the narrative sometimes lacks clarity and could benefit from a more focused scope. The vast amount of data and the complex mechanisms presented can be overwhelming and occasionally detract from the main message.

Methodological Concerns

The reliance on bioinformatics analyses and previously published datasets raises questions about the direct experimental validation of the proposed mechanisms. Although the computational approach is valid, the absence of direct experimental evidence or validation in breast cancer samples limits the strength of the conclusions.

Interpretation of Data

The interpretation of viral homology and its impact on cancer development is speculative in several sections. The connections made between EBV infections, chromosomal breakpoints, and cancerous mutations rely heavily on correlative data without sufficient causal evidence. A more cautious interpretation of the results, highlighting the need for further experimental validation, would strengthen the manuscript.

Consideration of Alternate Hypotheses

The paper could benefit from a more balanced discussion of alternative hypotheses explaining the observed data. For instance, the role of other environmental, genetic, or lifestyle factors in breast cancer development is not adequately considered. Acknowledging and discussing these potential confounders would provide a more comprehensive understanding of the complex etiology of breast cancer.

References and Current Literature

While the paper cites a significant amount of relevant literature, it sometimes overlooks recent studies that could either support or challenge the proposed hypotheses. Incorporating a more current and diverse range of references would enhance the paper's relevance and credibility.

Recommendation for Discussion Inclusion

To broaden the discussion and contextualize the findings within the broader research landscape, it is recommended to include the following article in the discussion section.

Al-Awaida W, Al-Ameer HJ, Sharab A, Akasheh RT. Modulation of wheatgrass (*Triticum aestivum Linn*) toxicity against breast cancer cell lines by simulated microgravity. Curr Res Toxicol. Sep 19, 2023;5:100127. [doi: 10.1016/j.crtox.2023.100127] [Medline: 37767028]

Incorporating this article could provide valuable insights into innovative approaches for studying cancer therapies. Specifically, the effects of simulated microgravity on the efficacy of natural compounds like wheatgrass against breast

cancer could open up new avenues for research on the environmental and physical conditions affecting cancer treatment outcomes. Discussing this study would enrich the manuscript by introducing the concept of microgravity as a novel factor influencing cancer cell behavior and therapy resistance, thereby offering a broader perspective on cancer research methodologies and therapeutic strategies.

Round 2 Review

General Comments

This paper tests the idea that EBV infections can help cause breast cancer by weakening the body's defenses against cancer. The study uses bioinformatics to compare chromosome breakpoints in breast cancer to those in cancers known to be caused by EBV. The results show that EBV might play a role in breast cancer by damaging important cell functions.

Specific Comments

Major Comments

The methods section needs more details about how the datasets were chosen and combined.

The discussion should explain more about how EBV might cause the chromosome breaks and rearrangements seen in breast cancer.

More data or references are needed to support the idea that EBV helps breast cancer spread to other parts of the body.

Minor Comments

Adding more references would strengthen the sections that talk about how EBV affects breast cancer.

Figures and tables should be clearly mentioned in the text to help readers follow the data.

Some parts of the manuscript need clearer writing and better organization, especially where complex bioinformatics results are explained.

The abstract should be revised to clearly highlight the main findings and why they are important.

Make sure all abbreviations are defined when they are first used to help readers understand the text better.

Conflicts of Interest

None declared.

Reference

1. Friedenson B. Identifying safeguards disabled by Epstein-Barr virus infections in genomes from patients with breast cancer: chromosomal bioinformatics analysis. JMIRx Med 2025;6:e50712. [doi: 10.2196/50712]

Abbreviations

EBV: Epstein-Barr virus

Edited by A Schwartz; submitted 13.12.24; this is a non–peer-reviewed article;accepted 13.12.24; published 29.01.25.

Please cite as:
Anonymous
Peer Review of "Identifying Safeguards Disabled by Epstein-Barr Virus Infections in Genomes From Patients With Breast Cancer: Chromosomal Bioinformatics Analysis"
JMIRx Med 2025;6:e70041
URL: https://xmed.jmir.org/2025/1/e70041
doi:10.2196/70041

Peer Review of “Mothers’ Knowledge of and Practices Toward Oral Hygiene of Children Aged 5-9 Years in Bangladesh: Cross-Sectional Study”

Bilkisu Nwankwo, MSc

Department of Community Medicine, College of Medicine, Kaduna State University, Kaduna, Nigeria

Related Articles:

Companion article: https://www.medrxiv.org/content/10.1101/2024.04.05.24305403v1

Companion article: https://med.jmirx.org/2025/1/e70145

Companion article: https://med.jmirx.org/2025/1/e59379

(JMIRx Med 2025;6:e70142) doi:10.2196/70142

KEYWORDS

mother's knowledge and practices; oral hygiene; child oral health; bangladesh

This is the peer-review report for “Mothers’ Knowledge of and Practices Toward Oral Hygiene of Children Aged 5-9 Years in Bangladesh: Cross-Sectional Study.”

Round 1 Review

Specific Comments

There were a lot of grammatical issues and typographical errors. The manuscript [1] needs to be edited for grammar and syntax. It is also obvious that the manuscript was not proofread adequately.

Major Comments

Abstract

- A word is missing in the first sentence. Authors should proofread the manuscript.
- Keywords: Dhaka is a more appropriate keyword than Bangladesh.
- Under the Results in the abstract, respondents should be referred to as such and not as samples.

Introduction

- The global prevalence of oral diseases was stated, but authors did not capture the prevalence in the study area/country and so have not shown that oral disease is a problem. Even the global prevalence that was stated was only that of dental caries among the seven conditions that make up oral diseases as stated by the authors.
- The objective stated here (last sentence) comes off like the authors are assessing the knowledge and practices of oral hygiene with regard to themselves and not their children as stated in the topic.

Methods

- Was it permission that was given by the institutional review board or an ethical clearance?
- This section is quite disorganized. There is a logical flow expected in this section.
- Why was a nonprobability sampling technique (convenient sampling) used for this study? The sampling technique was not explained at all. This will make replicating this study difficult.
- I have an issue with the scoring system and the grading. Is there a reference for it? I particularly have an issue with “moderately average.” It is not a standard term.
- The exclusion criteria are not the opposite of the inclusion criteria as stated by the authors. Exclusion criteria are those already included in the study but that are ineligible for one reason or the other.

Results

- In the text above Table 1, authors wrote that most respondents (39.3%) had a monthly family income of “21,000 - 40,000 taka per month.” This figure (39.3%) is just over one-third of the respondents and not a majority.
- Table 1: What is the meaning of graduation and above? Is it graduated secondary school or graduated college?
- “Respectively” should be added at the end of the following sentence. “Out of 400 mothers, more than 90% knew the importance of brushing teeth while 82.3% and 80.8% of them knew the recommended frequency and appropriate time for brushing teeth.”

Discussion

- The second sentence: the study is not evaluating parent’s knowledge and practices but that of mothers.
- Grammatical errors and missing words

Reference List

- Some of the references were not cited correctly. Authors should adhere to the Vancouver referencing style.

Conflicts of Interest

None declared.

Reference

1. Tamannur T, Das SK, Nesa A, et al. Mothers' knowledge of and practices toward oral hygiene of children aged 5-9 years in Bangladesh: cross-sectional study. JMIRx Med 2025;6:e59379. [doi: 10.2196/59379]

Edited by T Leung; submitted 16.12.24; this is a non–peer-reviewed article;accepted 06.01.25; published 03.02.25.

Please cite as:
Nwankwo B
Peer Review of "Mothers' Knowledge of and Practices Toward Oral Hygiene of Children Aged 5-9 Years in Bangladesh: Cross-Sectional Study"
JMIRx Med 2025;6:e70142
URL: https://xmed.jmir.org/2025/1/e70142
doi:10.2196/70142

Peer Review of “Mothers’ Knowledge of and Practices Toward Oral Hygiene of Children Aged 5-9 Years in Bangladesh: Cross-Sectional Study”

Md Hafizul Islam

Institute of Nutrition and Food Science, University of Dhaka, Dhaka, Bangladesh

Related Articles:

Companion article: https://www.medrxiv.org/content/10.1101/2024.04.05.24305403v1

Companion article: https://med.jmirx.org/2025/1/e70145

Companion article: https://med.jmirx.org/2025/1/e59379

(JMIRx Med 2025;6:e70144) doi:10.2196/70144

KEYWORDS

mothers’ knowledge and practices; oral hygiene; child oral health; Bangladesh

This is the peer-review report for “Mothers’ Knowledge of and Practices Toward Oral Hygiene of Children Aged 5-9 Years in Bangladesh: Cross-Sectional Study.”

Round 1 Review

This is an interesting piece of research [1], which highlights mothers’ knowledge and practices regarding their children’s oral health in Dhaka City. However, several issues made the study scientifically questionable. The major issues are as follows. The study included mothers from two hospitals in Dhaka City, but the title of the study does not mention this. The sample selection from the mothers visiting the hospitals might not represent general mothers from the whole of Dhaka. Thus, this study might not be generalizable to all mothers in Dhaka City.

Introduction

Revise the last paragraph of the Introduction to highlight the study gap in Bangladesh and clearly state the objective of the study. Use the formal word “mother” and avoid the word “moms.”

Methods

Study Setting and Participants

Give clear reasoning as to why you selected study participants from the hospitals. The last line is confusing. It is not clear whether the participants filled out the questionnaire on their own or they were interviewed by the enumerators.

Sampling Technique

Please mention the nonresponse bias for the convenient sampling. Give a short description of the pretesting mentioning the number of samples, period, and location for it.

Measurement of Knowledge and Practice Score

Give the 15 knowledge-related questions and 13 practice-related questions in the supplementary file. Mention if these questions are your own or if you used any valid tools or questions adopted from the relevant previous studies. Give adequate information regarding the scoring system of the variables, mentioning the highest possible aggregated score and examples of two questions (one for knowledge and one for practice).

Statistical Analyses

The authors mentioned that they used the Mann-Whitney *U* test and the Kruskal-Wallis test. However, they did not mention the underlying assumptions of the tests. Moreover, the Results section also shows the χ^2 test but is not mentioned in the Methods section. Furthermore, the last line of the Results of the abstract shows the Pearson correlation coefficient, but nothing is mentioned in the Methods or Results section of the entire manuscript.

Results

Table 1

It is confusing as the text description of Table 1 and the title of Table 1 are different. It is recommended to use two separate tables: one for socioeconomic variables and another for the frequency distribution of the knowledge level among socioeconomic variables. Mention the knowledge- and practice-related raw scores first and then the cross-tab results. There is a major mistake in the results of Tables 1 and 2. The frequency distribution for educational status, occupation, family type, number of family members, and monthly income in Tables 1 and 2 are the same. However, the *P* values are different. How is this possible? Please check the results.

Discussion

It is confusing whether the practice was for the children or how a mother takes care of their children's dental health. Mention the implications of your findings rather than just comparing the findings with previous studies. State the limitation of the study, especially the bias regarding convenient sampling. Provide a section on the public health significance of the study findings in Bangladesh.

Conclusion

The Conclusion section of the study is poorly written and not focused on the findings of the study. Revise the Conclusion section to highlight your study findings.

Round 2 Review

The authors impressively amended the initial version of the manuscript based on the reviewers' comments. However, several issues remain unaddressed.

1. The authors should include the city in the title of the study. You can revise the title to "Knowledge and practices towards oral hygiene of children aged 5 - 9 years old: a cross-sectional study among mothers visited tertiary level hospitals in Dhaka, Bangladesh."
2. Use the full form when it appears first and then use the abbreviation afterward. For example, "KP" in the abstract.
3. Please mention this statistical test in the Methods section of the abstract. You did not mention the χ^2 test and Pearson correlation.
4. It is recommended to make the recommendation simple and easy to understand for the readers. Avoid duplication of the same term.
5. In the sample size calculation, you used *P*=.58 and *P*=.57. Please clarify why you used those prevalences. Cite the relevant study here.
6. Before the heading for the sociodemographic variables in the Methods section, you mention outcome measures. However, the sociodemographic variables are not your outcome variables according to your objectives. You can remove the term outcome measures from here.
7. You mentioned that you used 13 questions for the assessment of practices. Thus, according to your scoring approach, there should be a score of 1-13, but here, it is 1-11.
8. Please mention the name of the software and version you used for the statistical analysis.
9. Revise the sentence before Table 1. You can make it two sentences. One for family income and another for occupation.
10. There is no chi-square–related data in Table 1. Please remove the footnotes from Table 1.
11. In Figure 1, it is recommended to keep the values to one decimal point for 1a and 1b.
12. Please revise the sentence before Table 3 to give a clear meaning.
13. You can remove the percentage symbol from the value and give it in the vertical axis title.
14. Please give the correlation results in the main manuscript or as a supplementary table.
15. The authors overlooked the association of knowledge and practice with income and family size. Please give more details on those two points in the Discussion section.

Conflicts of Interest

None declared.

Reference

1. Tamannur T, Das SK, Nesa A, et al. Mothers' knowledge of and practices toward oral hygiene of children aged 5-9 years in Bangladesh: cross-sectional study. JMIRx Med 2025;6:e59379. [doi: 10.2196/59379]

Edited by T Leung; submitted 16.12.24; this is a non–peer-reviewed article;accepted 16.12.24; published 03.02.25.

Please cite as:
Islam MH
Peer Review of "Mothers' Knowledge of and Practices Toward Oral Hygiene of Children Aged 5-9 Years in Bangladesh: Cross-Sectional Study"
JMIRx Med 2025;6:e70144
URL: https://xmed.jmir.org/2025/1/e70144
doi:10.2196/70144

Peer Review of “Converting Organic Municipal Solid Waste Into Volatile Fatty Acids and Biogas: Experimental Pilot and Batch Studies With Statistical Analysis”

Anonymous

Related Articles:

Companion article: https://www.biorxiv.org/content/10.1101/2023.06.21.545938v2

Companion article: https://med.jmirx.org/2025/1/e69894

Companion article: https://med.jmirx.org/2025/1/e50458

(JMIRx Med 2025;6:e69895) doi:10.2196/69895

KEYWORDS

multistep fermentation; specific methane production; anaerobic digestion; kinetics study; biochar; first-order; modified Gompertz; mass balance; waste management; environment sustainability

This is a peer-review report for “Converting Organic Municipal Solid Waste Into Volatile Fatty Acids and Biogas: Experimental Pilot and Batch Studies With Statistical Analysis”

Round 1 Review

The present manuscript [1] deals with the study of the valorization of organic fractions of municipal solid waste through the production of volatile fatty acids and biogas. The article is interesting; in my opinion, it should be revised.

Comments

1. The presentation of the manuscript is very poor; the figures are not in the same format.
2. Some of the recent works should be discussed and cited in the Introduction section: [2-6].
3. The novelty of the work should be highlighted.
4. Full stops should be removed from all subheadings.
5. The Results and Discussion should be written in detail with proper subheadings.
6. There are some typo errors; they should be rectified.

Conflicts of Interest

None declared.

References

1. Borhany H. Converting organic municipal solid waste into volatile fatty acids and biogas: experimental pilot and batch studies with statistical analysis. JMIRx Med 2025;6:e50458. [doi: 10.2196/50458]
2. Jung S, Shetti NP, Reddy KR, et al. Synthesis of different biofuels from livestock waste materials and their potential as sustainable feedstocks – a review. Energy Conversion Manage 2021 May 15;236:114038. [doi: 10.1016/j.enconman.2021.114038]
3. Srivastava RK, Shetti NP, Reddy KR, Aminabhavi TM. Sustainable energy from waste organic matters via efficient microbial processes. Sci Total Environ 2020 Jun 20;722:137927. [doi: 10.1016/j.scitotenv.2020.137927] [Medline: 32208271]
4. Sampath P, Brijesh, Reddy KR, et al. Biohydrogen production from organic waste – a review. Chem Eng Technol 2020 Jul;43(7):1240-1248. [doi: 10.1002/ceat.201900400]
5. Velvizhi G, Goswami C, Shetti NP, Ahmad E, Kishore Pant K, Aminabhavi TM. Valorisation of lignocellulosic biomass to value-added products: paving the pathway towards low-carbon footprint. Fuel (Lond) 2022 Apr 1;313:122678. [doi: 10.1016/j.fuel.2021.122678]
6. Monga D, Shetti NP, Basu S, et al. Engineered biochar: a way forward to environmental remediation. Fuel (Lond) 2022 Mar 1;311:122510. [doi: 10.1016/j.fuel.2021.122510]

Edited by T Leung; submitted 10.12.24; this is a non–peer-reviewed article;accepted 10.12.24; published 04.02.25.

Please cite as:
Anonymous
Peer Review of "Converting Organic Municipal Solid Waste Into Volatile Fatty Acids and Biogas: Experimental Pilot and Batch Studies With Statistical Analysis"
JMIRx Med 2025;6:e69895
URL: https://xmed.jmir.org/2025/1/e69895
doi:10.2196/69895

Peer Review of "Converting Organic Municipal Solid Waste Into Volatile Fatty Acids and Biogas: Experimental Pilot and Batch Studies With Statistical Analysis"

Dina Elsalamony, MSc

Department of Biotechnology, Institute of Graduate Studies & Research, University of Alexandria, Alexandria, Egypt

Related Articles:

Companion article: https://www.biorxiv.org/content/10.1101/2023.06.21.545938v2

Companion article: https://med.jmirx.org/2025/1/e69894

Companion article: https://med.jmirx.org/2025/1/e50458

(JMIRx Med 2025;6:e69896) doi:10.2196/69896

KEYWORDS

multistep fermentation; specific methane production; anaerobic digestion; kinetics study; biochar; first-order; modified Gompertz; mass balance; waste management; environment sustainability

This is a peer-review report for "Converting Organic Municipal Solid Waste Into Volatile Fatty Acids and Biogas: Experimental Pilot and Batch Studies With Statistical Analysis."

Round 1 Review

General Comments

Generally, the manuscript [1] should be strictly improved in English language writing and corrected for all grammatical errors throughout the whole manuscript. The author has to use a uniform style of the English language, either American or British English. Further English assistance is particularly required. Many missing articles and a lot of grammatical and punctuation errors must be corrected in the manuscript as in the corrected abstract.

Specific Comments

This paper shows an important aspect of multiple fermentation steps for the complete utilization of municipal solid waste and conversion to useful products, which is highly recommended for circular economic sustainability worldwide. However, it needs some major revision and arrangement to allow for a better presentation of this valuable work.

Major Comments

Title

1. "Valorization of Organic Fraction of Municipal Solid Waste Through Production of Volatile Fatty Acids (VFAs) and Biogas" is a long title that should be shortened to be more concise with no abbreviations—more indicative. Suggested title: "Valorization of Organic Municipal Solid Waste for Volatile Fatty Acids and Biogas Production."

Abstract Section

2. Generally speaking, it must be more concise and specific.

3. Please clearly mention the take-home message and the main findings of the research.

4. The abstract is too long and lacks the main methodology and main experimental techniques that were carried out in this work. The author may add some hints about the main methods used before mentioning the main results.

Manuscript

5. Keywords: Words must be modified to be more informative and representative of the research interest and differ from the word in the manuscript title. Maybe add "Multi Step of Fermentation Process" or "Waste Management and Environment Sustainability."

6. Arrangement of the experimental work in the manuscript may be needed in the Results and Discussion accordingly.

7. There is a lack of figures to describe the main parameter optimization steps well. Please reformulate to describe some data using figures with error bars.

8. The SD and table footnotes with the number of replicates should be noted underneath all of the given tables.

9. A mechanistic in-detail discussion is required, not just comparing your results with the previous work; justify better.

10. In research articles, do not include any table comparing literature results; the author can discuss the main findings in the text itself, as in Table 5.

11. The Conclusions section is missing in the manuscript to summarize and point out the novelty and the main findings from the research.

12. Generally speaking, in academic writing, (1) abstracts do not include abbreviations, (2) avoid articles in the title (the, a, an), and (3) avoid keywords that exist in the title.

13. As a rule of thumb, no dots in titles or subtitles as in the Experimental section, Anerobic Pilot Unities, etc.

14. Multiple references should be merged, not written separately, as in “29, 30” and “23, 27”; the author may use the merge reference option in reference software.

15. The author may add numbers for all titles and subtitles accordingly all over the manuscript.

Minor Comments

16. The author should avoid general and well-known information, and be selective in the recent references used. May add one small paragraph to the Biological Waste Management and Environment Sustainability section.

17. The author should clarify the main aim of the work clearly in the last paragraph of the Introduction.

18. Do not use our, we, or us in academic writing.

19. The author may mention novel applications of VFA and biogas. Mention different novel sources of biogas production.

20. The author should mention the gas chromatography type, gas injection rate, column dimensions, and the used carrier gas in the main document.

21. The author did not mention that flushing with nitrogen or carbon dioxide took place in anaerobic digestion while feeding reactors and how the anaerobic conditions were maintained; please mention it clearly or add the references used for the methodology.

22. Organize titles all over the manuscript.

23. Generally, the subtitles are too generic; modify them to be more indicative and precise.

24. “unless Saturday and Sunday” in line 208 is not important information; the suggested word “daily” is enough.

25. “Unite”: Please correct.

26. Remove the grid lines in the figures.

27. The author has to mention the range used for the chemical oxygen demand method, and the original reference should be cited appropriately.

28. “As can be seen”: This statement is repetitive more than once in the Discussion, in lines 301, 315, and 423.

29. Figure 3 caption: mesophilic fermentation: Please specify which stage because both of the sequential steps were called mesophilic fermentation in Figure 1.

30. What is the rationale for comparing 3 days to 4.5 days for all the used systems; the author may justify why 4.5 days is better to complete with this hydraulic retention time in the rest of the experiments or describe the one variable at a time optimization method that is used to determine the significant factors and the insignificant one; mention them clearly. Also, use in the Discussion the terms “significant” and “insignificant” according to the obtained *P* value.

31. The author has to mention tables and figures in the text in their appropriate place.

Round 2 Review

This paper is greatly enhanced compared to the previous copy, and the author followed the previous comments precisely.

I recommend its publication. Thanks for allowing me to review this interesting work.

General Note

The Word file is the correct revised one, but the PDF seems to be the old version.

Conflicts of Interest

None declared.

Reference

1. Borhany H. Converting organic municipal solid waste into volatile fatty acids and biogas: experimental pilot and batch studies with statistical analysis. JMIRx Med 2025;6:e50458. [doi: 10.2196/50458]

Abbreviations

VFA: volatile fatty acid

Edited by T Leung; submitted 10.12.24; this is a non–peer-reviewed article;accepted 10.12.24; published 04.02.25.

Please cite as:
Elsalamony D
Peer Review of "Converting Organic Municipal Solid Waste Into Volatile Fatty Acids and Biogas: Experimental Pilot and Batch Studies With Statistical Analysis"
JMIRx Med 2025;6:e69896
URL: https://xmed.jmir.org/2025/1/e69896
doi:10.2196/69896

Peer Review of “Determinants of Periodic Health Examination Uptake: Insights From a Jordanian Cross-Sectional Study”

Ali Ahmed, MPhil, PhD, PharmD

Division of Infectious Diseases and Global Public Health, School of Medicine, University of California, San Diego, La Jolla, CA, United States

Related Articles:

Companion article: https://www.medrxiv.org/content/10.1101/2024.02.03.24302286v1

Companion article: https://med.jmirx.org/2025/1/e71528

Companion article: https://med.jmirx.org/2025/1/e57597

(JMIRx Med 2025;6:e71529) doi:10.2196/71529

KEYWORDS

periodic health examination; PHE; preventive health services; routine health checkups; Jordan; cross-sectional study

This is a peer-review report for “Determinants of Periodic Health Examination Uptake: Insights From a Jordanian Cross-Sectional Study.”

Round 1 Review

Specific Comments

Major Comments

1. In this manuscript [1], write in detail about the data collection procedure.
2. Why was a convenience sampling technique employed?
3. “All collected data are treated with strict confidentiality.” Some language corrections are required.

Minor Comments

There are a lot of formatting issues; many things seem copied and pasted.

Conflicts of Interest

None declared.

Reference

1. Tayoun AA. Determinants of periodic health examination uptake: insights from a Jordanian cross-sectional study. JMIRx Med 2025;6:e57597. [doi: 10.2196/57597]

Edited by T Leung; submitted 20.01.25; this is a non–peer-reviewed article;accepted 20.01.25; published 05.02.25.

Please cite as:
Ahmed A
Peer Review of “Determinants of Periodic Health Examination Uptake: Insights From a Jordanian Cross-Sectional Study”
JMIRx Med 2025;6:e71529
URL: https://xmed.jmir.org/2025/1/e71529
doi:10.2196/71529

Peer Review of "Determinants of Periodic Health Examination Uptake: Insights From a Jordanian Cross-Sectional Study"

Anonymous

Related Articles:

Companion article: https://www.medrxiv.org/content/10.1101/2024.02.03.24302286v1

Companion article: https://med.jmirx.org/2025/1/e71528

Companion article: https://med.jmirx.org/2025/1/e57597

(JMIRx Med 2025;6:e71531) doi:10.2196/71531

KEYWORDS

periodic health examination; PHE; preventive health services; routine health checkups; Jordan; cross-sectional study

This is a peer-review report for "Determinants of Periodic Health Examination Uptake: Insights From a Jordanian Cross-Sectional Study."

Round 1 Review

The following items were noted in this paper [1].

- Periodic health examination (PHE) uptake: Only 27.1% of participants underwent a PHE in the last 2 years.
- Predictors: Significant predictors include recent visits to a primary health care facility, monthly income, and knowledge about PHEs and preventive health measures.
- Nonsignificant factors: Gender, marital status, smoking status, and BMI did not show a significant association with PHE uptake.

Strengths

1. Comprehensive analysis: The study employs a robust methodology, combining descriptive, inferential, and multivariate statistical techniques to provide a thorough understanding of PHE uptake.
2. Significant predictors identified: Key factors influencing PHE uptake were identified, offering valuable insights for health care providers and policy makers.
3. First of its kind in Jordan: This study fills a gap in existing knowledge by being the first to investigate PHE uptake in Jordan.

Negative Points and Areas for Improvement

Cross-Sectional Design

- Limitation: The study's design limits the ability to establish causality.
- Improvement: Future research could benefit from a longitudinal approach to better establish causal relationships between the identified predictors and PHE uptake.

Convenience Sampling

- Limitation: This method may introduce selection bias, and the online survey format may lead to measurement bias.
- Improvement: Employing a more randomized and stratified sampling method could enhance the representativeness and validity of the findings.

Limited Generalizability

- Limitation: Results may not be generalizable to populations outside of Jordan or those not included in the sample.
- Improvement: Expanding the study to include diverse populations and different geographic regions would provide a more comprehensive understanding of PHE uptake.

Survey Instrument

- Limitation: The questionnaire's comprehensiveness and relevance to the Jordanian context might not have been fully ensured.
- Improvement: Pretesting the survey with a larger and more varied group, followed by adjustments based on feedback, could improve its applicability and accuracy.

Behavioral Factors

- Limitation: The study did not find a relationship between behavioral factors and PHE uptake, which contradicts findings in other contexts.
- Improvement: A more detailed investigation into cultural and societal influences on health behaviors in Jordan is needed to clarify these results.

English Language and Clarity

- Limitation: The manuscript contains some grammatical errors and awkward phrasings, which can detract from its readability.
- Improvement: A thorough review and editing for language and clarity by a native English speaker or professional editor would enhance the manuscript's quality.

Conflicts of Interest

None declared.

Reference

1. Tayoun AA. Determinants of periodic health examination uptake: insights from a Jordanian cross-sectional study. JMIRx Med 2025;6:e57597. [doi: 10.2196/57597]

Abbreviations

PHE: periodic health examination

Edited by T Leung; submitted 20.01.25; this is a non–peer-reviewed article;accepted 20.01.25; published 05.02.25.

Please cite as:
Anonymous
Peer Review of "Determinants of Periodic Health Examination Uptake: Insights From a Jordanian Cross-Sectional Study"
JMIRx Med 2025;6:e71531
URL: https://xmed.jmir.org/2025/1/e71531
doi:10.2196/71531

Peer Review for “Checklist Approach to Developing and Implementing AI in Clinical Settings: Instrument Development Study”

Anonymous

Related Articles:

Companion article: https://www.medrxiv.org/content/10.1101/2024.08.08.24311701v1

Companion article: https://med.jmirx.org/2025/1/e69537

Companion article: https://med.jmirx.org/2025/1/e65565

(JMIRx Med 2025;6:e69870) doi:10.2196/69870

KEYWORDS

artificial intelligence; machine learning; algorithm; model; analytics; AI deployment; human-AI interaction; AI integration; checklist; clinical workflow; clinical setting; literature review

This is the peer-review report for “Checklist Approach to Developing and Implementing AI in Clinical Settings: Instrument Development Study.”

Round 1 Review

General Comments

This paper [1] construct a checklist to support the development and implementation of artificial intelligence (AI) in clinical settings. I only have some minor comments.

Minor Comments

1. Comparison with existing checklists: Please add a comparison with some of the existing checklists.

2. Inconsistency in the number of studies: The authors initially stated that they included 20 studies, but later mentioned 23. Please clarify.

3. Appendix visibility: The appendix is not visible.

4. Abbreviation consistency: The abbreviation “IQR” appears multiple times. Please ensure it is clearly defined and used consistently.

Conflicts of Interest

None declared.

Reference

1. Owoyemi A, Osuchukwu J, Salwei ME, Boyd A. Checklist approach to developing and implementing AI in clinical settings: instrument development study. JMIRx Med 2025;6:e65565. [doi: 10.2196/65565]

Abbreviations

AI: artificial intelligence

Edited by CN Hang, E Meinert, T Leung; submitted 10.12.24; this is a non–peer-reviewed article;accepted 10.12.24; published 20.02.25.

Please cite as:
Anonymous
Peer Review for “Checklist Approach to Developing and Implementing AI in Clinical Settings: Instrument Development Study”
JMIRx Med 2025;6:e69870
URL: https://xmed.jmir.org/2025/1/e69870
doi:10.2196/69870

Peer Review for “Checklist Approach to Developing and Implementing AI in Clinical Settings: Instrument Development Study”

Saima Zaki

Department of Physiotherapy, School of Allied Health Sciences, Sharda University, Plot No. 32-34, Knowledge Park III, Greater Noida, India

Related Articles:

Companion article: https://www.medrxiv.org/content/10.1101/2024.08.08.24311701v1

Companion article: https://med.jmirx.org/2025/1/e69537

Companion article: https://med.jmirx.org/2025/1/e65565

(JMIRx Med 2025;6:e70058) doi:10.2196/70058

KEYWORDS

artificial intelligence; machine learning; algorithm; model; analytics; AI deployment; human-AI interaction; AI integration; checklist; clinical workflow; clinical setting; literature review

This is the peer-review report for “Checklist Approach to Developing and Implementing AI in Clinical Settings: Instrument Development Study.”

Round 1 Review

General Comments

This paper [1] presents the Clinical Artificial Intelligence (AI) Sociotechnical Framework (CASoF), a checklist intended to support the development and implementation of AI systems in health care settings. The framework is built on a comprehensive literature review and a modified Delphi study involving health care professionals globally. The manuscript addresses a significant gap in the integration of AI by emphasizing the importance of sociotechnical considerations alongside technical aspects.

Specific Comments

Major Comments

1. Clarity and structure: The manuscript could benefit from clearer explanations, particularly in the methodology section. The description of the Delphi study and literature synthesis is dense and may be difficult for readers to follow. Consider breaking down complex sentences and using more straightforward language.

2. Methodological rigor: The manuscript lacks details on the selection process for Delphi panelists and the exact methods used for data analysis. Transparency in these areas would significantly strengthen the paper. Additionally, clarify how the Delphi method was modified and the rationale behind these modifications.

3. Literature review and contextualization: The discussion section could benefit from a more critical comparison between the CASoF and existing frameworks. While the manuscript mentions other frameworks, it does not fully explore their limitations or how the CASoF overcomes these challenges. Expanding this discussion would provide a stronger justification for the CASoF’s novelty and utility.

4. Checklist practicality: While the checklist is comprehensive, its length and complexity may hinder practical adoption. Consider providing a condensed version for quick reference and include examples of how the checklist can be applied in real-world scenarios.

5. Ethical considerations and bias mitigation: The manuscript discusses ethical considerations but lacks specific strategies for addressing these issues within the CASoF. Expanding this discussion would enhance the manuscript’s comprehensiveness.

Minor Comments

6. Typographical and grammatical errors: There are minor typographical and grammatical errors throughout the manuscript that should be corrected. For instance, phrases like “revised and edited” could be simplified to “revised” for conciseness.

7. Tables and figures formatting: The tables summarizing the Delphi study results are helpful but could be more effectively formatted. Using shading or color coding to distinguish between different stages or domains would improve clarity and ease of interpretation.

8. Recent references: Some references in the manuscript are relatively old, which is less ideal for a rapidly evolving field like AI. Where possible, the manuscript should incorporate more recent literature to support its claims and demonstrate the ongoing relevance of the topic.

Conflicts of Interest

None declared.

Reference

1. Owoyemi A, Osuchukwu J, Salwei ME, Boyd A. Checklist approach to developing and implementing AI in clinical settings: instrument development study. JMIRx Med 2025;6:e65565. [doi: 10.2196/65565]

Abbreviations

AI: artificial intelligence
CASoF: Clinical Artificial Intelligence Sociotechnical Framework

Edited by CN Hang, E Meinert, T Leung; submitted 13.12.24; this is a non–peer-reviewed article;accepted 13.12.24; published 20.02.25.

Please cite as:
Zaki S
Peer Review for "Checklist Approach to Developing and Implementing AI in Clinical Settings: Instrument Development Study"
JMIRx Med 2025;6:e70058
URL: https://xmed.jmir.org/2025/1/e70058
doi:10.2196/70058

Peer Review for “Checklist Approach to Developing and Implementing AI in Clinical Settings: Instrument Development Study”

Anonymous

Related Articles:

Companion article: https://www.medrxiv.org/content/10.1101/2024.08.08.24311701v1

Companion article: https://med.jmirx.org/2025/1/e69537

Companion article: https://med.jmirx.org/2025/1/e65565

(JMIRx Med 2025;6:e69869) doi:10.2196/69869

KEYWORDS

artificial intelligence; machine learning; algorithm; models; analytics; AI deployment; human-AI interaction; AI integration; checklist; clinical workflow; clinical setting; literature review

This is the peer-review report for “Checklist Approach to Developing and Implementing AI in Clinical Settings: Instrument Development Study.”

Round 1 Review

The paper [1] presents the Clinical Artificial Intelligence (AI) Sociotechnical Framework (CASoF), a structured approach to guide the planning, design, development, and implementation of AI systems in health care settings. The framework is designed to address the gap between technical performance and sociotechnical factors that are essential for successful AI deployment in clinical environments.

The authors conducted a literature synthesis and a modified Delphi study involving global health care professionals to develop and refine the CASoF checklist. The checklist emphasizes the importance of considering the value proposition, data integrity, human-AI interaction, technical architecture, organizational culture, and ongoing support and monitoring, to ensure that AI tools are not only technologically sound but also practically viable and socially adaptable within clinical settings.

The study found that the successful integration of AI in health care depends on a balanced focus on both technological advancements and the sociotechnical environment of clinical settings. The CASoF represents a step forward in bridging this divide, offering a holistic approach to AI deployment that is mindful of the complexities of health care systems. The checklist aims to facilitate the development of AI tools that are effective, user-friendly, and seamlessly integrated into clinical workflows, ultimately enhancing patient care and health care outcomes.

The authors acknowledge some limitations of the study, such as the need for continuous refinement of the CASoF through iterative feedback and broader engagement with more stakeholders. Future research should aim to include an even wider array of perspectives, particularly from underrepresented regions and specialties, to enhance the framework’s comprehensiveness and applicability.

Overall, the paper provides a valuable contribution to the field of AI in health care by offering a practical and comprehensive approach to the development and implementation of AI systems in clinical settings.

Conflicts of Interest

None declared.

Reference

1. Owoyemi A, Osuchukwu J, Salwei ME, Boyd A. Checklist approach to developing and implementing AI in clinical settings: instrument development study. JMIRx Med 2025;6:e65565. [doi: 10.2196/65565]

Abbreviations

AI: artificial intelligence
CASoF: Clinical Artificial Intelligence Sociotechnical Framework

Edited by CN Hang, E Meinert, T Leung; submitted 10.12.24; this is a non–peer-reviewed article;accepted 10.12.24; published 20.02.25.

Please cite as:
Anonymous
Peer Review for "Checklist Approach to Developing and Implementing AI in Clinical Settings: Instrument Development Study"
JMIRx Med 2025;6:e69869
URL: https://xmed.jmir.org/2025/1/e69869
doi:10.2196/69869

Peer Review for “Checklist Approach to Developing and Implementing AI in Clinical Settings: Instrument Development Study”

Keith Thompson, MD

Department of Family Medicine, Western University, 1151 Richmond St, London, Canada

Related Articles:

Companion article: https://www.medrxiv.org/content/10.1101/2024.08.08.24311701v1

Companion article: https://med.jmirx.org/2025/1/e69537

Companion article: https://med.jmirx.org/2025/1/e65565

(JMIRx Med 2025;6:e69593) doi:10.2196/69593

KEYWORDS

artificial intelligence; machine learning; algorithm; model; analytics; AI deployment; human-AI interaction; AI integration; checklist; clinical workflow; clinical setting; literature review

This is the peer-review report for “Checklist Approach to Developing and Implementing AI in Clinical Settings: Instrument Development Study.”

Round 1 Review

General Comments

This paper [1]...is a very cohesive approach to establishing a framework for the implementation of artificial intelligence (AI).

Specific Comments

Major Comments

1. Ideally there should be information on the demographics of the expert panel.

2. Please add comments regarding equitable access for these technologies.

Conflicts of Interest

None declared.

Reference

1. Owoyemi A, Osuchukwu J, Salwei ME, Boyd A. Checklist approach to developing and implementing AI in clinical settings: instrument development study. JMIRx Med 2025;6:e65565. [doi: 10.2196/65565]

Abbreviations

AI: artificial intelligence

Edited by CN Hang, E Meinert, T Leung; submitted 03.12.24; this is a non–peer-reviewed article;accepted 03.12.24; published 20.02.25.

Please cite as:
Thompson K
Peer Review for “Checklist Approach to Developing and Implementing AI in Clinical Settings: Instrument Development Study”
JMIRx Med 2025;6:e69593
URL: https://xmed.jmir.org/2025/1/e69593
doi:10.2196/69593

Peer Review for "Checklist Approach to Developing and Implementing AI in Clinical Settings: Instrument Development Study"

Sai Saripalli, MSc

Louisiana State University, Baton Rouge, LA, United States

Related Articles:

Companion article: https://www.medrxiv.org/content/10.1101/2024.08.08.24311701v1

Companion article: https://med.jmirx.org/2025/1/e69537

Companion article: https://med.jmirx.org/2025/1/e65565

(JMIRx Med 2025;6:e69594) doi:10.2196/69594

KEYWORDS

artificial intelligence; machine learning; ML; algorithm; model; analytics; AI deployment; human-AI interaction; AI integration; checklist; clinical workflow; clinical setting; literature review

This is the peer-review report for "Checklist Approach to Developing and Implementing AI in Clinical Settings: Instrument Development Study."

Round 1 Review

General Comments

Using artificial intelligence (AI) to add social and domain-specific steps to clinical trials is innovative [1]. My only comment is whether the number of stages or the checklist changes if the shortlisted panelists change.

Specific Comments

Major Comments

1. I am unsure if having 38 (expert) panelists is good enough to have a robust framework.

Conflicts of Interest

None declared.

Reference

1. Owoyemi A, Osuchukwu J, Salwei ME, Boyd A. Checklist approach to developing and implementing AI in clinical settings: instrument development study. JMIRx Med 2025;6:e65565. [doi: 10.2196/65565]

Abbreviations

AI: artificial intelligence

Edited by CN Hang, E Meinert, T Leung; submitted 03.12.24; this is a non–peer-reviewed article;accepted 03.12.24; published 20.02.25.

Please cite as:
Saripalli S
Peer Review for "Checklist Approach to Developing and Implementing AI in Clinical Settings: Instrument Development Study"
JMIRx Med 2025;6:e69594
URL: https://xmed.jmir.org/2025/1/e69594
doi:10.2196/69594

Peer Review for “Checklist Approach to Developing and Implementing AI in Clinical Settings: Instrument Development Study”

Anonymous

Related Articles:

Companion article: https://www.medrxiv.org/content/10.1101/2024.08.08.24311701v1

Companion article: https://med.jmirx.org/2025/1/e69537

Companion article: https://med.jmirx.org/2025/1/e65565

(JMIRx Med 2025;6:e69595) doi:10.2196/69595

KEYWORDS

artificial intelligence; machine learning; algorithm; model; analytics; AI deployment; human-AI interaction; AI integration; checklist; clinical workflow; clinical setting; literature review

This is the peer-review report for “Checklist Approach to Developing and Implementing AI in Clinical Settings: Instrument Development Study.”

Round 1 Review

This paper [1] introduces the Clinical Artificial Intelligence (AI) Sociotechnical Framework (CASoF), a checklist developed through a literature synthesis and refined by a modified Delphi study. It aims to guide the development and implementation of AI in clinical settings, focusing on the integration of both technological performance and sociotechnical factors. The framework addresses gaps in existing frameworks by emphasizing not only technical specifications but also the broader sociotechnical dynamics essential for successful AI deployment in health care.

New approaches to reporting AI in clinical settings are crucial as AI becomes more integrated into clinical practice. However, the paper needs to address the “black box” dilemma more thoroughly. This refers to the opaque nature of AI algorithms, where the decision-making process is not easily interpretable by clinicians, leading to trust and transparency issues. Additionally, while the CASoF checklist is a valuable tool, it would benefit from a more detailed comparison to established frameworks like TRIPOD (Transparent Reporting of a Multivariable Prediction Model for individual Prognosis or Diagnosis), which has been widely used in developing and validating clinical prediction models. Discussing how the CASoF complements or improves upon TRIPOD would strengthen the paper’s contributions.

I suggest adding a paragraph discussing the potential roles of AI when integrated into hospital electronic health record (EHR) systems. AI could be used for the development of advanced diagnostic and prognostic tools by analyzing real-time patient data. Integration with EHRs could enhance decision-making, providing predictive analytics at the point of care and improving patient outcomes. This would help explore the broader clinical impact of AI beyond just technical integration, addressing its potential for continuous learning and optimization in health care settings.

Conflicts of Interest

None declared.

Reference

1. Owoyemi A, Osuchukwu J, Salwei ME, Boyd A. Checklist approach to developing and implementing AI in clinical settings: instrument development study. JMIRx Med 2025;6:e65565. [doi: 10.2196/65565]

Abbreviations

AI: artificial intelligence
CASoF: Clinical Artificial Intelligence Sociotechnical Framework
EHR: electronic health record
TRIPOD: Transparent Reporting of a Multivariable Prediction Model for individual Prognosis or Diagnosis

Edited by CN Hang, E Meinert, T Leung; submitted 03.12.24; this is a non–peer-reviewed article;accepted 03.12.24; published 20.02.25.

Please cite as:
Anonymous
Peer Review for "Checklist Approach to Developing and Implementing AI in Clinical Settings: Instrument Development Study"
JMIRx Med 2025;6:e69595
URL: https://xmed.jmir.org/2025/1/e69595
doi:10.2196/69595

Peer Review for “Impact of a Point-of-Care Ultrasound Training Program on the Management of Patients With Acute Respiratory or Circulatory Failure by In-Training Emergency Department Residents (IMPULSE): Before-and-After Implementation Study”

Anonymous

Related Articles:

Companion article: https://www.medrxiv.org/content/10.1101/2023.09.28.23295699v1

Companion article: https://med.jmirx.org/2025/1/e72092

Companion article: https://med.jmirx.org/2025/1/e53276

(JMIRx Med 2025;6:e72144) doi:10.2196/72144

KEYWORDS

point-of-care ultrasonography; training program; acute respiratory failure; acute circulatory failure; emergency department

This is the peer-review report for “Impact of a Point-of-Care Ultrasound Training Program on the Management of Patients With Acute Respiratory or Circulatory Failure by In-Training Emergency Department Residents (IMPULSE): Before-and-After Implementation Study.”

Round 1 Review

General Comments

This paper [1] researches an essential component of point-of care ultrasonography. As this modality is rapidly evolving, evaluation of the impact on patient management and outcomes as well as cost-effectiveness is essential. Both subjects discussed in the paper result in a highly relevant manuscript. Even though the authors discuss relevant subjects, there are some problems with the manuscript.

Specific Comments

Major Comments

1. The title of the manuscript suggests that the authors researched the impact of ultrasound after implementation. However, as stated in the Methods section, ultrasound is already used by senior physicians. Thus, the impact of ultrasound after implementation is not researched but rather the impact of ultrasound used by residents. I suggest that the authors clarify that this is a feasibility and impact study on the implementation of point-of-care ultrasound (POCUS) used by residents in the emergency department (ED) in the title and Abstract.

2. The authors state that patients were not included consecutively due to logistics in phase 2. This results in a high risk of bias in the included patients. Please include in the CONSORT (Consolidated Standards of Reporting Trials) diagram the number of patients that were eligible and were excluded based on exclusion criteria or missed.

3. It is unclear how many residents were performing the ultrasound examinations included in the analysis: the Methods section state that there was only 1 resident at the ED in both phases, while in the Results section, it states that there were 12 residents trained. Please clarify.

4. The authors state that they chose a before-and-after implementation to evaluate the effect of POCUS to avoid contamination. However, before the implementation, POCUS was already used by senior physicians, which raises the question if POCUS was indeed not used in phase 1 of the trial.

5. Interestingly, in the Discussion section, the author discussed that the publication of Msolli et al [2] did not find an improvement of diagnostic accuracy. It would be interesting to discuss why this is the case.

6. In the Discussion and Conclusion, it is suggested that the use of POCUS might lead to a decrease in hospital mortality. Since this is an observational study in which, just as the authors state, a diagnostic tool rather than a therapeutic intervention is researched, this is too rash to state. Please remove this from the Conclusion and Abstract.

Minor Comments

Overall

7. The authors provide results with IQR; however, no ranges are given. Please describe results as mean (SD) when data are normally distributed or median (25th percentile – 75th percentile) when data are not normally distributed.

8. Formatting of the full manuscript needs some attention. For example, in the Abstract, not all sentences start with a capital letter. Also, it is common in the English language to write number in full up to 20.

9. Please follow the author guidelines of the journal for reporting values and the structure of the manuscript.

Title Page

10. The authors state that a clinical trial registration was done. However, it seems that they refer to a registration by a medical ethical review board. Please provide a clinical trial registration or if not applicable, remove it from the title page.

Introduction

11. In the first sentence, please state the full name of "emergency department" before using the abbreviation ED.

Methods

12. Figure 1 should be formatted. The most common formatting is according to the CONSORT flow diagram.

Results

13. Please do not discuss the results in the Results section.

Discussion

14. Please end the Discussion section with the strengths and limitations. The secondary findings should be above the Strengths and Limitations section.

Round 2 Review

I would like to compliment the authors of their extensive changes to the manuscript. I have some minor comments.

Minor Comments

1. I would suggest changing the sentence "However, there is still no strong evidence that the diagnostic accuracy of POCUS translates into a clinically relevant difference in patient outcomes" in the Introduction, because you also do not provide strong evidence (I do not know if we ever could provide strong evidence). I would suggest that you focus it more on the fact that the impact of using POCUS in the management of patients in the ED is still relatively unknown.

2. I would suggest to start the Discussion section with a short summary of the key findings.

Conflicts of Interest

None declared.

References

1. Bieler S, Tagan D, Grosgurin O, Fumeaux T. Impact of a point-of-care ultrasound training program on the management of patients with acute respiratory or circulatory failure by in-training emergency department residents (IMPULSE): before-and-after implementation study. JMIRx Med 2025;6:e53276. [doi: 10.2196/53276]
2. Msolli MA, Sekma A, Marzouk MB, et al. Bedside lung ultrasonography by emergency department residents as an aid for identifying heart failure in patients with acute dyspnea after a 2-h training course. Ultrasound J 2021 Feb 9;13(1):5. [doi: 10.1186/s13089-021-00207-9] [Medline: 33559777]

Abbreviations

CONSORT: Consolidated Standards of Reporting Trials
ED: emergency department
POCUS: point-of-care ultrasound

Edited by E Meinert, A Schwartz; submitted 04.02.25; this is a non–peer-reviewed article;accepted 04.02.25; published 03.03.25.

Please cite as:
Anonymous
Peer Review for "Impact of a Point-of-Care Ultrasound Training Program on the Management of Patients With Acute Respiratory or Circulatory Failure by In-Training Emergency Department Residents (IMPULSE): Before-and-After Implementation Study"
JMIRx Med 2025;6:e72144
URL: https://xmed.jmir.org/2025/1/e72144
doi:10.2196/72144

Peer Review of "Predicting Escalation of Care for Childhood Pneumonia Using Machine Learning: Retrospective Analysis and Model Development"

Colin Rogerson, MD, MPH

Division of Pediatric Critical Care, Regenstrief Center for Biomedical Informatics, Indiana University School of Medicine, 705 Riley Hospital Drive, Indianapolis, IN, United States

Related Articles:

Companion article: https://www.medrxiv.org/content/10.1101/2024.02.22.24303209v1

Companion article: https://med.jmirx.org/2025/1/e71098

Companion article: https://med.jmirx.org/2025/1/e57719

(JMIRx Med 2025;6:e71100) doi:10.2196/71100

KEYWORDS

childhood pneumonia; community-acquired pneumonia; machine learning; clinical decision support system; prognostic care decision

This is the peer-review report for "Predicting Escalation of Care for Childhood Pneumonia Using Machine Learning: Retrospective Analysis and Model Development."

Round 1 Review

General Comments

The authors [1] have examined the medical records for 437 patients with pneumonia and created a machine learning–based classifier to determine which patients required transfer to a tertiary care center. This subject is interesting, as the predictive power of these novel statistical techniques is high and could improve the clinical care of these patients. The authors have done thorough work describing the statistical methods used in the preprocessing of the data and model development. My primary concerns in the manuscript are the lack of clinical application description, the lack of description of the time frame of the included data elements, and the lack of description regarding the patient population and outcome of interest. The following are my point-by-point comments.

Specific Comments

Major Comments

Abstract

- The authors use the term "case management" in the Abstract and several times in the manuscript. In this context, the authors' meaning is the decision for the escalation of care or patient transfer. However, in US-based hospital systems, case management has a different meaning, which includes largely transition to rehabilitation or nursing facilities, acquisition of home oxygen therapy, etc. I would recommend altering this term for comprehension to something like "escalation of care" or "patient triage."
- The primary outcome of interest should be included in the Abstract.
- As detailed in the Methods section, it is crucial to describe the time frame for the included variables, to know when the algorithm could be used in clinical practice.

Introduction

- As the goal of the algorithm in the study is to predict which patients will need transfer to tertiary care for increasing respiratory support, more of the Introduction should focus on the management of in-hospital pediatric pneumonia, challenges, and reasons for the escalation of care.
- I would recommend altering the sentence that describes pneumonia as easily preventable and treatable. Several of the most complicated cases in the intensive care unit are admitted with pneumonia.

Methods

- While great care is taken to describe the approach to data preprocessing, feature selection, and model development, I would recommend following the TRIPOD (Transparent Reporting of a Multivariable Prediction Model for individual Prognosis or Diagnosis) guidelines [2], which are validated reporting recommendations for predictive models.
- Please provide more details regarding the hospital systems involved in this study. Are they large, academic centers or small, rural centers?
- For study inclusion, I am not familiar with the Integrated Management of Childhood Illness guidelines. Are these structured diagnostic codes captured in the electronic health record? Is it a computational phenotype?

- Please specify what is meant by “neonatal age.”
- Many of the variables included in the model are colinear. For example, age and weight are highly dependent on one another, and including both in the model can be detrimental. The feature selection methods may be able to discern this, but maybe not. I would recommend using only age and *z* score in the model.
- The time frames are not stated for the variables. For example, does “hypoxia” mean hypoxia at any time during the hospitalization? On hospital admission? In the first 12 hours? This information is vital to determine the usability of the entire model. If the model uses variables available during the entire hospitalization, the predictive ability will be high, but the usability will be low. A model that can predict right when a patient is transferred to a tertiary care center that the patient will be transferred is useless. However, a model that can predict on admission, or in the first 6 - 12 hours, that a patient will require transfer is incredibly helpful. Without knowing the time frame for these variables, we cannot assess how the model could be applied in clinical practice.
- Please provide clarity regarding the study outcomes. The primary outcome is described as whether the patient was referred to a tertiary care center or not. The next sentence describes “poor prognosis” as pediatric intensive care unit admission or oxygen/ventilation support. How is this outcome used? Is this a secondary outcome? Is this describing the reason for transfer? Please clarify.
- As stated in the TRIPOD guidelines, you should present the amount of missingness in your data. It appears you used imputation methods for missing data. It is helpful to describe the amount of missing data that was imputed and the method for imputation.

Results

- There is a glaring lack of information regarding your study population. Please provide a table describing patient characteristics including demographics and the variables you used in the algorithm. Also, please provide a comparison between the patients who were transferred to a tertiary care center and those who were not.
- In imbalanced datasets, it can be more useful to measure model performance using the area under the precision-recall curve rather than the standard area under the receiver operator characteristic curve. I would recommend adding this metric.

Discussion

- The Discussion, overall, focuses much more on the technical details of the data curation and model development than it does on the clinical application of the model. Much of the technical details presented are also clearly explained in the Methods section and then repeated in the Discussion. I would recommend substantial revision to the Discussion section to remove redundant information that is already contained in the Methods section, as well as the addition of how this model could be applied in a clinical setting to improve the care of patients with pneumonia.
- The Discussion contains no information regarding the limitations of the study. Please describe in detail the prominent limitations of the study. These should include the use of retrospective data, including only two centers, imbalanced data, challenges with clinical implementation of the model, etc.
- The Discussion, and other areas of the manuscript, mention disease prevention several times. The goal of this study has nothing to do with the prevention of pneumonia, only the treatment of pneumonia and the prevention of associated morbidity and mortality. Please revise.

Conclusion

- As it stands, the Conclusion is fairly long and does not focus only on the primary findings of the study. I would recommend trimming it to 2 - 3 sentences that focus only on the primary findings of the study, such as the feasibility of developing this type of predictive model and the potential applications of the model to clinical practice.

Minor Comments

Methods

- The authors describe that ensemble methods “significantly enhance the accuracy of classifications.” Please provide a reference for this statement.

Results

- Please provide numbers for those who met your primary outcome of interest (transfer to a tertiary care center).
- Please provide a description of the time frame for patient transfer, for those who were transferred.

Discussion

- It would be interesting to hear more regarding the use of this model in resource-limited settings and the benefits it could provide.

Round 2 Review

General Comments

The authors have conducted a single-center, retrospective study evaluating the derivation and performance of a machine learning model to predict the need for transfer to a higher level of care for childhood pneumonia. The authors were provided with a substantial amount of feedback on the original submission, and although the authors’ response is detailed and comments on how all concerns were adequately addressed, the resulting manuscript is lacking in many if not most of the requested changes. The revised manuscript remains confusing to the reader and bereft of some essential elements of standard study reporting, including a basic description of the patient population and details regarding the timing of variable collection and use in the model. Due to this lack of response to the initial reviewer feedback, I am recommending rejection of this manuscript. The following are my point-by-point critiques, many of which are similar to those in my original review.

Specific Comments

Abstract

- First sentence: Please revise it to "Pneumonia is the leading cause of preventable mortality for children under five years of age."
- Background: The terms "case management" and "disease prevention" are still used in the Abstract. In my initial review, I recommended revising these terms to improve study clarity, and although the authors stated in their response that they replaced these terms, they remain in the Abstract. As it stands, it is not immediately clear to the reader that the purpose of the study was to provide a tool to assist bedside clinicians to determine which patients are likely to require transfer of care to a higher-level facility for pediatric pneumonia.
- Methods: As it stands, it is confusing to the readers what was actually done in the study. It should be very apparent that the authors used a specific list of variables (please provide each in the Abstract) to predict the need for transfer to a larger institution using a specific type of machine learning model (ensemble). In the current version, this is difficult to discern.
- Results: I would be completely clear regarding the outcome your model is predicting. After reading the paper, it is understood that "pneumonia prognosis" and "severity" actually mean required transfer to a higher level of care, but it is unclear in the Abstract. I would explicitly state "predicted transfer to a higher level of care with 77% - 88% accuracy."

Introduction

- Second paragraph, fifth sentence: I would recommend revising it to "However, this preventable health problem continues to be a substantial cause of mortality, especially in underdeveloped countries and regions, due to the lack of equipment and trained human resources." There is no way to quantify it as "the most important cause of mortality."
- The term "case management" continues to be used in the Introduction, which decreases clarity for the reader.
- As recommended previously, I would be very specific in the Introduction that you are trying to create a tool to help bedside clinicians (typically non–intensive care physicians) decide when to transfer a patient with pneumonia to a higher level of care to prevent morbidity and mortality. As it stands, this is unclear.

Methods

- In my initial review, I asked the authors to clarify what is meant by neonatal age. In their response, they said they had revised the Methods to state specifically 28 days or fewer. However, in the first paragraph of the Methods, it continues to state "neonatal age." Please revise.
- For clarity, I would recommend restating your primary outcome to simply "required tertiary care referral." Having the outcome as severe versus nonsevere, which is defined as requiring tertiary care referral or not, adds an extra step to the thought process and can be confusing.
- One of my largest concerns in the initial manuscript was the timing of the variables. This is crucial when determining how useful the model could be. If the elements in Table 1 are measured on admission, or in the first 6 - 12 hours of admission, the model could be very useful for patient care. If the elements were measured at any point during the hospitalization, it becomes much less useful. My worry is that the model was developed based on the elements' presence at any point, meaning if the child had fever, cough, respiratory distress, and hypoxia at hour 48, then at hour 49 the model was able to predict the patient would need transfer, and the patient was transferred at hour 50—this is not helpful to clinicians. On the other hand, if the model predicts at hour 12 that a patient needs transfer, and then at hour 50 they transfer, that is potentially very helpful to clinicians. Without these details, I cannot recommend the publication of the manuscript.
- It appears that the model was developed using the data from all 437 patients, and the results are presented following k-fold cross validation. It is standard practice to derive the model on a subset of the data (typically 70% - 80%) and then to test it on the remainder of the dataset to prevent overfitting and inflation of performance metrics. It does not appear that this was done. Despite having a small sample size, I believe this approach would lead to a more robust and generalizable model.

Results

- The first paragraph contains many "nuts and bolts" details of model development, and these would be better positioned in the Methods section.
- Both reviewers on the initial submission requested additional details describing the study population, and although the authors responded that they added these details, there are still none provided. It is essential to the understanding of the study results to know the characteristics of the patient population, and it should be a standard requirement for all clinical studies.
- The Shapley additive explanations value results presented in Figure 2 are valuable, but more details describing each measured factor are required. I recommend a table with each factor as rows and two columns comparing the population that did not require transfer to a tertiary care center to the population that did.
- An additional figure showing an area under the precision-recall curve for each model would also be interesting to the readers.

Discussion

- The Discussion spends a decent amount of space discussing the COVID-19 pandemic. While this does have some bearing on the management of childhood pneumonia, I believe the space would be better spent discussing the actual implementation of this type of algorithm. How would a primary care clinician actually use this model in practice? How would it improve upon current clinical practice? Would it be easy or difficult to incorporate into routine workflows? This would be more interesting to the readers.

- I recommend adding what the next steps of this line of research would be. How would you seek to improve the model's performance? More patient data? Additional variables?
- In the original submission, I recommended the authors provide a limitations section and also provided some examples. Although the authors response says they added this, there are still no limitations provided. Please provide this essential element to the Discussion.

Conclusion

- I recommend commenting on what the next steps of this line of research would be in more specific terms.

Round 3 Review

General Comments

The authors have conducted a single-center, retrospective study evaluating the derivation and performance of a machine learning model to predict the need for transfer to a higher level of care for childhood pneumonia. The authors were provided with a substantial amount of feedback on the original submission and have been responsive to feedback, which has resulted in a much improved manuscript. There remain several typographical and grammatical errors, which I would advise an English-grammar expert to review prior to publication, but from a scientific standpoint, I believe the manuscript is appropriate for publication.

Specific Comments

Major Comments

1. Details regarding the patient population have been provided in detail.

2. The study objectives have been clarified for readers.

3. The study methods are now much more reproducible.

Conflicts of Interest

None declared.

References

1. Serin O, Akbasli IT, Cetin SB, et al. Predicting escalation of care for childhood pneumonia using machine learning: retrospective analysis and model development. JMIRx Med 2025;e57719:6. [doi: 10.2196/57719]
2. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent Reporting of a Multivariable Prediction Model for individual Prognosis or Diagnosis (TRIPOD): the TRIPOD statement. BMJ 2015 Jan 7;350:g7594. [doi: 10.1136/bmj.g7594] [Medline: 25569120]

Abbreviations

TRIPOD: Transparent Reporting of a Multivariable Prediction Model for individual Prognosis or Diagnosis

Edited by E Meinert, S Amal, T Leung; submitted 09.01.25; this is a non–peer-reviewed article;accepted 09.01.25; published 04.03.25.

Please cite as:
Rogerson C
Peer Review of "Predicting Escalation of Care for Childhood Pneumonia Using Machine Learning: Retrospective Analysis and Model Development"
JMIRx Med 2025;6:e71100
URL: https://xmed.jmir.org/2025/1/e71100
doi:10.2196/71100

Peer Review of “Predicting Escalation of Care for Childhood Pneumonia Using Machine Learning: Retrospective Analysis and Model Development”

Anonymous

Related Articles:

Companion article: https://www.medrxiv.org/content/10.1101/2024.02.22.24303209v1

Companion article: https://med.jmirx.org/2025/1/e71098

Companion article: https://med.jmirx.org/2025/1/e57719

(JMIRx Med 2025;6:e71369) doi:10.2196/71369

KEYWORDS

childhood pneumonia; community-acquired pneumonia; machine learning; clinical decision support system; prognostic care decision

This is the peer-review report for “Predicting Escalation of Care for Childhood Pneumonia Using Machine Learning: Retrospective Analysis and Model Development.”

Round 1 Review

General Comments

This paper [1] developed a machine learning approach that could predict community-acquired pneumonia prognosis, which is scaled into two levels, severe or nonsevere, and identify important clinical indices, such as hypoxia, respiratory distress, age, *z* score of weight for age, and antibiotic usage before admission. The machine learning–based clinical decision support system tool for childhood pneumonia could provide prognostic support for case management.

Specific Comments

Major Comments

1. To enhance the manuscript’s grounding in current research and to provide a comprehensive context for the study, the authors are recommended to incorporate an evaluation of related literature in the Introduction and Discussion sections. This could include, but not be limited to, the following studies:

- Liu YC, Cheng HY, Chang TH, et al. Evaluation of the need for intensive care in children with pneumonia: machine learning approach. JMIR Med Inform. Jan 27, 2022;10(1):e28934. [doi: 10.2196/28934] [Medline: 35084358]
- Smith JC, Spann A, McCoy AB, et al. Natural language processing and machine learning to enable clinical decision support for treatment of pediatric pneumonia. AMIA Annu Symp Proc. Jan 25, 2020;2020:1130-1139. [Medline: 33936489]
- Kanwal K, Khalid SG, Asif M, Zafar F, Qurashi AG. Diagnosis of community-acquired pneumonia in children using photoplethysmography and machine learning-based classifier. Biomed Signal Process Control. Jan 2024;87:105367. [doi: 10.1016/j.bspc.2023.105367]
- Chang TH, Liu YC, Lin SR, et al. Clinical characteristics of hospitalized children with community-acquired pneumonia and respiratory infections: Using machine learning approaches to support pathogen prediction at admission. J Microbiol Immunol Infect. Aug 2023;56(4):772-781. [doi: 10.1016/j.jmii.2023.04.011] [Medline: 37246060]

The readers could have a more comprehensive understanding if the authors could include a concise evaluation of the prior literature in the current manuscript.

2. Considering the high stakes involved in pediatric care, particularly in intensive settings, it is critical to exam the false negative cases from the confusion matrices. Analyzing these cases for any common feature characteristics could provide insights into potential improvements in the predictive algorithm. This analysis should be clearly presented and discussed in the manuscript, emphasizing its importance in clinical decision-making.

3. The manuscript would benefit from a more detailed description of the cohort used in the study. Information on age, gender, and other clinical indices across the two groups (severe and nonsevere) would enable a better understanding of the study population. Additionally, providing the number of cases in each group would clarify the scope and scale of the study findings.

4. A detailed description of the data collection process is crucial for assessing the study’s applicability in real-world clinical settings. The manuscript should explicitly state the following:

- How and when clinical data, including features such as hypoxia and respiratory distress, were collected (eg, at the time of admission? or within 24 hours of admission?);

- The time frame considered for “antibiotic usage before admission” as relevant to the prediction model: This information is essential for replicability and for future applications of the findings in clinical workflows.

Round 2 Review

I thank the authors for revising the manuscript.

Conflicts of Interest

None declared

Reference

1. Serin O, Akbasli IT, Cetin SB, et al. Predicting escalation of care for childhood pneumonia using machine learning: retrospective analysis and model development. JMIRx Med 2025;6:e57719. [doi: 10.2196/57719]

Edited by E Meinert; submitted 16.01.25; this is a non–peer-reviewed article;accepted 16.01.25; published 04.03.25.

Please cite as:
Anonymous
Peer Review of “Predicting Escalation of Care for Childhood Pneumonia Using Machine Learning: Retrospective Analysis and Model Development”
JMIRx Med 2025;6:e71369
URL: https://xmed.jmir.org/2025/1/e71369
doi:10.2196/71369

Peer Review of "Data Obfuscation Through Latent Space Projection for Privacy-Preserving AI Governance: Case Studies in Medical Diagnosis and Finance Fraud Detection"

Reenu Singh

Indian Institute of Management Mumbai, Mumbai, India

Related Articles:

Companion article: https://arxiv.org/abs/2410.17459v1

Companion article: https://med.jmirx.org/2025/1/e72527

Companion article: https://med.jmirx.org/2025/1/e70100

(JMIRx Med 2025;6:e72523) doi:10.2196/72523

KEYWORDS

privacy-preserving AI; latent space projection; data obfuscation; AI governance; machine learning privacy; differential privacy; k-anonymity; HIPAA; GDPR; compliance; data utility; privacy-utility trade-off; responsible AI; medical imaging privacy; secure data sharing; LSP; artificial intelligence

This is a peer-review report for "Data Obfuscation Through Latent Space Projection for Privacy-Preserving AI Governance: Case Studies in Medical Diagnosis and Finance Fraud Detection."

Round 1 Review

Specific Comments

Major Comments

1. What was the basis of taking up health care cancer diagnosis and financial fraud for the study [1]? Will latent space projection be an effective method for privacy protection in speech therapy to analyze audio datasets to assist in diagnosing and treating speech-related disorders; in medical imaging video datasets from endoscopy, ultrasounds, and robotic surgeries for diagnostics and artificial intelligence–assisted tools; and in telemedicine to analyze video feeds for remote consultations and diagnoses?
2. The basic structure of the paper is missing. Please follow the guidelines of journal paper writing with distinctly visible sections of Introduction, Method, Result/Findings, Discussion, and Limitations with future scope and conclusion. The introduction, background, and related work should be written cohesively, and all should come under the Introduction heading.
3. The statistical tables are in excess. The tables and values should be talked about in written form. Limit the number of images and tables to 5 - 6 or according to the journal guidelines. Use an appendix for the flowchart and any other tabular data that is too lengthy.
4. Explanations of tables and figures should be in paragraph form. Please cite literature where comparative inference and process-specific benefits and drawbacks are mentioned. Examples are Tables 1-5. For writing sections like "Comparative Analysis with Existing Techniques," all the subparts should be written in paragraphs and discuss the values and analysis only, and put them in their respective paragraphs, removing the tabular data. Please use appendices for excessive tables. Within the body of the research paper, 5 - 6 figures and tables are sufficient; the rest should be put in appendices.
5. In "Latency and Performance analysis, part A" and "Performance optimization" are mentions of the literature, which should be present as part of the literature in the Introduction paragraph. Restating the literature again is redundant. Stick to the structure of the journal paper. Please cite references to support the claims, such as "real-time requirements of financial systems" under the section of Real-Time Performance.
6. "Scalability analysis" and other sections: What were the criteria for the choice of datasets for the study for the case studies? What were the data sizes? Give specifications in the first paragraph of respective case studies. Presenting the details about the process of procurement of files, data extraction, limitations in data handling, etc. Are there any limitations in adopting the latent space projection methods?

Round 2 Review

General Comments

This paper is highly relevant to health care, particularly in the context of privacy management of data during the analysis of imagery.

Specific Comments

Major Comments

1. The case studies should be written in a more descriptive style. Please reduce the use of numbered or bullet points (in the Introduction, Method, and Result) to align with the formal writing style typically suitable for journal papers.

2. Please rephrase the description of Table 3 (immediately following the table) in a narrative style. This approach enhances the readability of the article.

3. Two figures should not be positioned consecutively. Include some text between Figure 3 and Figure 4. Adjust and reorganize the content to ensure a smooth flow.

Minor Comments

4. The titles of tables and figures should be presented as captions. Revise the captions to ensure they do not begin with a verb.

Conflicts of Interest

None declared.

Reference

1. Vaijainthymala Krishnamoorthy M. Data Obfuscation Through Latent Space Projection for Privacy-Preserving AI Governance: Case Studies in Medical Diagnosis and Finance Fraud Detection. JMIRx Med 2025;6:e70100. [doi: 10.2196/70100]

Edited by CN Hang; submitted 11.02.25; this is a non–peer-reviewed article;accepted 11.02.25; published 12.03.25.

Please cite as:
Singh R
Peer Review of "Data Obfuscation Through Latent Space Projection for Privacy-Preserving AI Governance: Case Studies in Medical Diagnosis and Finance Fraud Detection"
JMIRx Med 2025;6:e72523
URL: https://xmed.jmir.org/2025/1/e72523
doi:10.2196/72523

Peer Review of “Data Obfuscation Through Latent Space Projection for Privacy-Preserving AI Governance: Case Studies in Medical Diagnosis and Finance Fraud Detection”

Trutz Bommhardt

University of Wuppertal, Wuppertal, Germany

Related Articles:

Companion article: https://arxiv.org/abs/2410.17459v1

Companion article: https://med.jmirx.org/2025/1/e72527

Companion article: https://med.jmirx.org/2025/1/e70100

(JMIRx Med 2025;6:e72525) doi:10.2196/72525

KEYWORDS

privacy-preserving AI; latent space projection; data obfuscation; AI Governance; machine learning privacy; differential privacy; k-anonymity; HIPAA; GDPR; compliance; data utility; privacy-utility trade-off; responsible AI; medical imaging privacy; secure data sharing; LSP; artificial intelligence

This is a peer-review report for “Data Obfuscation Through Latent Space Projection for Privacy-Preserving AI Governance: Case Studies in Medical Diagnosis and Finance Fraud Detection.”

Round 1 Review

General Comments

I thoroughly enjoyed reading this paper [1] as it is a well-written article that will make an important contribution to the literature on the development of privacy-preserving artificial intelligence (AI) governance. I have attached a few comments to improve the study.

Specific Comments

Major Comments

Something like a discussion that embeds the latent space projection for AI governance and the results in the current scientific debate is missing before or after Chapter VII.

Minor Comments

In Chapter II B (Existing privacy-preserving techniques), please provide some further sources to demonstrate that the challenges mentioned are still relevant, as some sources are relatively old (eg, from 2009).

Conflicts of Interest

None declared.

Reference

1. Vaijainthymala Krishnamoorthy M. Data Obfuscation Through Latent Space Projection for Privacy-Preserving AI Governance: Case Studies in Medical Diagnosis and Finance Fraud Detection. JMIRx Med 2025;6:e70100. [doi: 10.2196/70100]

Abbreviations

AI: artificial intelligence

Edited by CN Hang; submitted 11.02.25; this is a non–peer-reviewed article;accepted 11.02.25; published 12.03.25.

Please cite as:
Bommhardt T
Peer Review of "Data Obfuscation Through Latent Space Projection for Privacy-Preserving AI Governance: Case Studies in Medical Diagnosis and Finance Fraud Detection"
JMIRx Med 2025;6:e72525
URL: https://xmed.jmir.org/2025/1/e72525
doi:10.2196/72525

Peer Review of “Improved Alzheimer Disease Diagnosis With a Machine Learning Approach and Neuroimaging: Case Study Development”

Anonymous

Related Articles:

Companion article: https://arxiv.org/abs/2405.09553v1

Companion article: https://med.jmirx.org/2025/1/e72821

Companion article: https://med.jmirx.org/2025/1/e60866

(JMIRx Med 2025;6:e73768) doi:10.2196/73768

KEYWORDS

Alzheimer disease; computer-aided diagnosis system; machine learning; principal component analysis; linear discriminant analysis; t-distributed stochastic neighbor embedding; feedforward neural network; vision transformer architecture; support vector machines; magnetic resonance imaging; positron emission tomography imaging; Open Access Series of Imaging Studies; Alzheimer's Disease Neuroimaging Initiative; OASIS; ADNI

This is a peer-review report for “Improved Alzheimer Disease Diagnosis With a Machine Learning Approach and Neuroimaging: Case Study Development.”

Round 1 Review

General Comments

This paper [1] proposes a computer-aided diagnosis (CAD) system for Alzheimer disease (AD) using principal component analysis (PCA) and machine learning–based approaches. The authors claim that their system, which combines PCA for feature extraction with support vector machines (SVMs) and artificial neural networks (ANNs) for classification, achieves good accuracy in detecting AD from magnetic resonance imaging (MRI) and positron emission tomography (PET) images. However, the paper could be strengthened by addressing several areas for improvement.

Specific Comments

Major Comments

1. Consideration of alternative methodologies: While the use of PCA, SVMs, and ANNs for AD classification is a valid approach, the authors should consider exploring more recent deep learning architectures, such as vision transformers, which have demonstrated state-of-the-art performance in medical image analysis. This would help to situate the work within the broader context of current research in the field.
2. Limited evaluation: The evaluation is limited to the Open Access Series of Imaging Studies (OASIS) dataset, which may not be representative of the diverse AD population. The authors should evaluate their system on larger and more diverse datasets, such as the Alzheimer’s Disease Neuroimaging Initiative (ADNI) dataset, to demonstrate its generalizability.

Minor Comments

1. Insufficient implementation details: The implementation details of the SVMs and ANNs are insufficient. The authors should specify the hyperparameters used, such as the kernel type and regularization parameters for SVMs, and the number of layers and neurons for ANNs.
2. Limited discussion: The discussion of the results is limited. The authors should provide a more in-depth analysis of the performance of their system, comparing it with other state-of-the-art methods and discussing the limitations and potential future directions.
3. The authors should ensure consistent formatting throughout the paper, including the use of italics for variables and proper capitalization in section headings.
4. The paper could be improved by using more precise language. For instance, instead of “good accuracy,” the authors could specify the exact accuracy percentage achieved by their system.

Round 2 Review

General Comments

This paper investigates the performance of various machine learning models in the diagnosis of AD using neuroimaging data. The authors propose a CAD system that uses PCA for feature extraction and SVMs, feedforward neural networks, and vision transformers for classification. The models are trained and evaluated on two datasets, OASIS and ADNI.

Specific Comments

Major Comments

1. The paper claims that the proposed CAD system is effective in classifying patients with AD and healthy controls with high accuracy. However, the reported accuracies of 91.9% for OASIS and 88.6% for ADNI using PCA/SVM are not significantly higher than those achieved by existing state-of-the-art methods (eg, Li Y, Chen G, Wang G, et al. Dominating Alzheimer's disease diagnosis with deep learning on sMRI and DTI-MD. *Front Neurol*. Aug 15, 2024;15:1444795. [doi: 10.3389/fneur.2024.1444795] [PMID: 39211812]). A more comprehensive literature review and comparison are needed to support the claim of the proposed system's superiority.
2. The ADNI dataset includes not only patients with AD and healthy controls but also individuals with mild cognitive impairment (MCI). The paper does not explicitly mention whether MCI cases are included in the ADNI dataset used in this study and if patients with MCI are excluded. What is the reason?
3. The paper's conclusion that the "PCA/SVM scheme is much better at predicting AD than the other models" is not supported by the results presented. The vision transformer model with data augmentation consistently outperforms PCA/SVM in terms of accuracy and other metrics. There are no obvious reasons data augmentation is unwanted either.

Minor Comments

1. The paper claims to use a multimodal system, combining both MRI and PET images. However, it does not compare the multimodal system's performance against single-modal systems using only MRI or PET images. Such a comparison would help to rationalize the conclusion that the multimodal system truly improves upon single-modal systems.

Conflicts of Interest

None declared.

Reference

1. Lazli L. Improved Alzheimer Disease Diagnosis With a Machine Learning Approach and Neuroimaging: Case Study Development. JMIRx Med 2025;6:e60866. [doi: 10.2196/60866]

Abbreviations

AD: Alzheimer disease
ADNI: Alzheimer's Disease Neuroimaging Initiative
ANN: artificial neural network
CAD: computer-aided diagnosis
MCI: mild cognitive impairment
MRI: magnetic resonance imaging
OASIS: Open Access Series of Imaging Studies
PCA: principal component analysis
PET: positron emission tomography
SVM: support vector machine

Edited by CN Hang; submitted 11.03.25; this is a non–peer-reviewed article;accepted 11.03.25; published 21.04.25.

Please cite as:
Anonymous
Peer Review of "Improved Alzheimer Disease Diagnosis With a Machine Learning Approach and Neuroimaging: Case Study Development"
JMIRx Med 2025;6:e73768
URL: https://xmed.jmir.org/2025/1/e73768
doi:10.2196/73768

Peer Review of "Improved Alzheimer Disease Diagnosis With a Machine Learning Approach and Neuroimaging: Case Study Development"

Masoud Khani

University of Wisconsin-Milwaukee, Milwaukee, WI, United States

Related Articles:

Companion article: https://arxiv.org/abs/2405.09553v1

Companion article: https://med.jmirx.org/2025/1/e72821

Companion article: https://med.jmirx.org/2025/1/e60866

(JMIRx Med 2025;6:e73454) doi:10.2196/73454

KEYWORDS

Alzheimer disease; computer-aided diagnosis system; machine learning; principal component analysis; linear discriminant analysis; t-distributed stochastic neighbor embedding; feedforward neural network; vision transformer architecture; support vector machines; magnetic resonance imaging; positron emission tomography imaging; Open Access Series of Imaging Studies; Alzheimer's Disease Neuroimaging Initiative; OASIS; ADNI

This is a peer-review report for "Improved Alzheimer Disease Diagnosis With a Machine Learning Approach and Neuroimaging: Case Study Development."

Round 1 Review

General Comments

This paper [1] explores the use of principal component analysis (PCA) and machine learning approaches for the diagnosis of Alzheimer disease (AD) using magnetic resonance imaging and positron emission tomography images from the Open Access Series of Imaging Studies database. The authors propose a system that combines PCA for feature extraction with artificial neural networks (ANNs) and support vector machines (SVMs) for classification. The paper is well structured and presents a clear methodology, but there are several areas where improvements are needed to enhance the rigor and impact of the research.

Specific Comments

Major Comments

1. Methodology justification: The choice of PCA as the sole feature extraction method needs further justification. While PCA effectively reduces dimensionality, it might not capture the most discriminative features of AD. Comparing PCA with other dimensionality reduction techniques like linear discriminant analysis or t-distributed stochastic neighbor emulation could provide a more comprehensive understanding of its effectiveness.
2. Evaluation metrics: The paper primarily focuses on accuracy as the evaluation metric. For medical diagnosis systems, metrics like sensitivity, specificity, precision, recall, and F_1-score are crucial as they provide a better understanding of the model's performance, especially in imbalanced datasets. Including these metrics would strengthen the evaluation section.
3. Dataset and preprocessing: The preprocessing steps are briefly mentioned but lack detailed explanation. Specific steps for noise reduction, intensity normalization, and any augmentation techniques used should be clearly described. Additionally, the impact of these preprocessing steps on the model's performance should be discussed.
4. Comparison with existing methods: The paper lacks a thorough comparison with existing state-of-the-art methods. Including a detailed comparison with recent literature, both in terms of methodology and performance, would provide better context and highlight the novelty and effectiveness of the proposed approach.

Minor Comments

1. Introduction section: The Introduction provides a good overview of AD and the need for early diagnosis. However, it could benefit from a more detailed discussion of the current challenges in AD diagnosis and how the proposed method aims to address these challenges.
2. Figure and table clarity: Figures and tables should be more clearly labeled and described. For example, in Table 1, it is unclear what "Total cost (Validation)" refers to. Additionally, the axes and legends in figures should be more descriptive to enhance readability.
3. Algorithm parameters: The specific parameters used for the SVMs and ANNs (eg, kernel type for SVMs, number of layers, and neurons for ANNs) should be explicitly

mentioned. This would help in reproducing the results and understanding the model configuration.

4. Conclusion and future work: The conclusion should be concise and focus on key findings. The Future Work section could be expanded to include more specific directions for further research, such as exploring different feature extraction methods, incorporating longitudinal data, or integrating other imaging modalities.
5. References: Ensure all references are up-to-date and relevant. Given the rapid advancements in machine learning and medical imaging, some references are slightly outdated. Including more recent studies would enhance the credibility and relevance of the paper.

Round 2 Review

General Comments

Thank you for addressing my comments from the previous round of reviews. I appreciate the effort you have put into revising the manuscript. The updated version effectively resolves all the issues I raised, and the manuscript is now clear, well-structured, and scientifically sound.

Conflicts of Interest

None declared.

Reference

1. Lazli L. Improved Alzheimer Disease Diagnosis With a Machine Learning Approach and Neuroimaging: Case Study Development. JMIRx Med 2025;6:e60866. [doi: 10.2196/60866]

Abbreviations

AD: Alzheimer disease
ANN: artificial neural network
PCA: principal component analysis
SVM: support vector machine

Edited by CN Hang; submitted 04.03.25; this is a non–peer-reviewed article;accepted 04.03.25; published 21.04.25.

Please cite as:
Khani M
Peer Review of "Improved Alzheimer Disease Diagnosis With a Machine Learning Approach and Neuroimaging: Case Study Development"
JMIRx Med 2025;6:e73454
URL: https://xmed.jmir.org/2025/1/e73454
doi:10.2196/73454

Peer Review for “Improved Alzheimer Disease Diagnosis With a Machine Learning Approach and Neuroimaging: Case Study Development”

Anonymous

Related Articles:

Companion article: https://arxiv.org/abs/2405.09553v1

Companion article: https://med.jmirx.org/2025/1/e72821

Companion article: https://med.jmirx.org/2025/1/e60866

(JMIRx Med 2025;6:e73130) doi:10.2196/73130

KEYWORDS

Alzheimer disease; computer-aided diagnosis system; machine learning; principal component analysis; linear discriminant analysis; t-distributed stochastic neighbor embedding; feedforward neural network; vision transformer architecture; support vector machines; magnetic resonance imaging; positron emission tomography imaging; Open Access Series of Imaging Studies; Alzheimer's Disease Neuroimaging Initiative; OASIS; ADNI

This is a peer-review report for “Improved Alzheimer Disease Diagnosis With a Machine Learning Approach and Neuroimaging: Case Study Development.”

Round 1 Review

General Comments

The paper [1] discusses the development of a machine learning–based computer-aided diagnosis system for the detection and classification of Alzheimer disease. The system uses brain magnetic resonance imaging and positron emission tomography images from the Open Access Series of Imaging Studies database, applying principal component analysis for feature extraction and using support vector machines (SVMs) and artificial neural networks (ANNs) as classifiers. Although the proposed model shows relatively good performance, the paper should focus on justifying the novelty of the method and providing more details in the results.

Specific Comments

Major Comments

1. The paper lacks a clear discussion on how the proposed method substantially advances the state of the art. While it combines principal component analysis with SVM and ANN, similar combinations have been explored in prior research. The authors should clearly write about how their work is novel and the specific contributions made beyond existing studies.
2. The paper does not provide sufficient details on the hyperparameter tuning process for both SVM and ANN models. The review suggests that the author include these additional details in an appendix.
3. The evaluation primarily focuses on accuracy, sensitivity, and specificity. However, other metrics like precision, F_1-score, and area under the receiver operating characteristic curve could provide a more comprehensive assessment of the model’s performance. The authors could consider adding additional metrics for evaluation.
4. In Figure 2, the size of the box on the left and right are different (square vs rectangle). Is there a specific reason the author made this design choice?

Minor Comments

1. The paper’s organization can be improved. Some sections, like the methodological explanation of principal component analysis, are overly detailed, while others, like the description of SVM and ANN, are relatively brief. Please consider balancing the sections.
2. The Related Work section is somewhat sparse and does not sufficiently cover recent advances in the field. Please consider adding more recent studies.

Conflicts of Interest

None declared.

Reference

1. Lazli L. Improved Alzheimer Disease Diagnosis With a Machine Learning Approach and Neuroimaging: Case Study Development. JMIRx Med 2025;6:e60866. [doi: 10.2196/60866]

Abbreviations

ANN: artificial neural network
SVM: support vector machine

Edited by CN Hang; submitted 25.02.25; this is a non–peer-reviewed article;accepted 25.02.25; published 21.04.25.

Please cite as:
Anonymous
Peer Review for "Improved Alzheimer Disease Diagnosis With a Machine Learning Approach and Neuroimaging: Case Study Development"
JMIRx Med 2025;6:e73130
URL: https://xmed.jmir.org/2025/1/e73130
doi:10.2196/73130

Peer Review of "Levels and Predictors of Knowledge, Attitudes, and Practices Regarding Contraception Among Female TV Studies Undergraduates in Nigeria: Cross-Sectional Study"

Kamal Kanti Biswas, MBBS, MBA

IPAS Bangladesh, House 428/A, Road 30 (3rd Floor), Dhaka, Bangladesh

Related Articles:

Companion article: https://www.medrxiv.org/content/10.1101/2024.01.01.24300698v1

Companion article: https://med.jmirx.org/2025/1/e72947

Companion article: https://med.jmirx.org/2025/1/e56135

(JMIRx Med 2025;6:e72949) doi:10.2196/72949

KEYWORDS

knowledge; attitudes; practice; contraception; regression; cross-sectional; females; students; Nigeria

This is a peer-review report for "Levels and Predictors of Knowledge, Attitudes, and Practices Regarding Contraception Among Female TV Studies Undergraduates in Nigeria: Cross-Sectional Study."

Round 1 Review

General Comments

Dear Authors,

Thank you very much for undertaking the study [1] titled "Levels and predictors of knowledge, attitude and practice of contraception among female TV undergraduates in Nigeria: a cross-sectional study" and submitting the manuscript to JMIR. The study findings are important for family planning program implementation targeting young students. I have the following comments and observations for improving your manuscript for consideration of publishing.

Specific Comments

Major Comments

Introduction: line 50: "youth": Indicate age group.

Line 52: "Utilization is higher": Not clear what the utilization was for.

Study population: limitation: gender biased. Male involvement and attitude are equally important regarding sexually transmitted infections, particularly for male methods like use of condoms. This needs to be mentioned as a limitation of the study.

Tables all: Hastily, one sentence is used for describing findings in a table. Need to elaborate more. Further comments below.

Table 1: Rephrase the "Marital status" indicator; the data does not give the status of marriage!

Table 2: Indicate what is meant by poor, good, etc, knowledge/attitude; cite measurement scale here.

Table 3: Need to mention if this was an open-ended or structured question.

Table 4: Cite the indicators used for measuring attitude toward use of contraception.

Table 5: The predictor of not engaging in sex may be reflected well in statistical analysis, but what is the significance in real life? Why would those who had never engaged in sex have used contraception?

Discussion: Mention the rate of use of emergency contraceptive pills (ECPs) also. This is increasing in many societies. Policy makers/planners are often not aware of the need for ECPs to include a supply of ECPs in a program. A recommendation like "There may be a need to use social marketing 42 approaches to make these contraceptives available to young people to bypass the stigma they experienced while accessing 43 contraceptives from traditional sources of contraceptives" is not supported by any finding or data of the study. Rather this raises a question of bias on jumping to a solution through a particular channel. Let the program planners find out the way to resolve the issue of information availability.

Highlights: Move the highlights to the Discussion section because this is a summary of the findings.

Conclusion: Rewrite the conclusion, elaborating on recommendations per the results of the study.

Conflicts of Interest

None declared.

Reference

1. Agbo HA, Adeoye PA, Yilzung DR, Mangut JS, Ogbada PF. Levels and predictors of knowledge, attitudes, and practices regarding contraception among female TV studies undergraduates in Nigeria: cross-sectional study. JMIRx Med 2025;6:e56135. [doi: 10.2196/56135]

Abbreviations

ECP: emergency contraceptive pill

Edited by A Schwartz; submitted 21.02.25; this is a non–peer-reviewed article;accepted 21.02.25; published 08.05.25.

Please cite as:
Biswas KK
Peer Review of "Levels and Predictors of Knowledge, Attitudes, and Practices Regarding Contraception Among Female TV Studies Undergraduates in Nigeria: Cross-Sectional Study"
JMIRx Med 2025;6:e72949
URL: https://xmed.jmir.org/2025/1/e72949
doi:10.2196/72949

Peer Review of "Levels and Predictors of Knowledge, Attitudes, and Practices Regarding Contraception Among Female TV Studies Undergraduates in Nigeria: Cross-Sectional Study"

Bilkisu Nwankwo, MBBS, MSc

Kaduna State University, Tafawa Balewa Way, Kaduna State, Nigeria

Related Articles:

Companion article: https://www.medrxiv.org/content/10.1101/2024.01.01.24300698v1

Companion article: https://med.jmirx.org/2025/1/e72947

Companion article: https://med.jmirx.org/2025/1/e56135

(JMIRx Med 2025;6:e72951) doi:10.2196/72951

KEYWORDS

knowledge; attitudes; practice; contraception; regression; cross-sectional; females; students; Nigeria

This is the peer-review report for "Levels and Predictors of Knowledge, Attitudes, and Practices Regarding Contraception Among Female TV Studies Undergraduates in Nigeria: Cross-Sectional Study."

Round 1 Review

Specific Comments

Major Comments

1. The sampling technique used in this paper [1] should be more detailed than it is. Respondents were said to have been selected by balloting from the 6 levels. Was it equal allocation per level, or was it proportionate allocation considering that it is not likely that there were the same number of students in each level?

2. State the age ranges of a teenager and that of a young adult in your methodology that informed the categorization in the Results.

3. Living with a spouse and not living with a spouse was considered for marital status in your study as opposed to being single, married, etc. Clarify why this is so.

4. The public health implications of some of the findings were omitted in the Discussion. This should be included. Its importance cannot be overemphasized.

Minor Comments

5. Abstract: The last sentence in the Methods is hanging. Kindly complete it.

6. Grammatical issues: Tenses: Future and present tenses were used where past tense should have been used in the methodology (lines 12 and 28). Present tense was used in multiple places in the Discussion where past tense should have been used.

7. Reference list: In the Vancouver referencing style, the month of publication should not appear as it did in some references like 7, 11, and 12. The date accessed/cited was written in some and not in others like 9, 10, 13, and 16. Really old references like reference 24, which is 14 years old, should be replaced by more current ones.

Conflicts of Interest

None declared.

Reference

1. Agbo HA, Adeoye PA, Yilzung DR, Mangut JS, Ogbada PF. Levels and predictors of knowledge, attitudes, and practices regarding contraception among female TV studies undergraduates in Nigeria: cross-sectional study. JMIRx Med 2025;6:e56135. [doi: 10.2196/56135]

Edited by A Schwartz; submitted 21.02.25; this is a non–peer-reviewed article;accepted 21.02.25; published 08.05.25.

Please cite as:
Nwankwo B
Peer Review of "Levels and Predictors of Knowledge, Attitudes, and Practices Regarding Contraception Among Female TV Studies Undergraduates in Nigeria: Cross-Sectional Study"
JMIRx Med 2025;6:e72951
URL: https://xmed.jmir.org/2025/1/e72951
doi:10.2196/72951

Peer Review of "Prevalence and Determinants of Academic Bullying Among Junior Doctors in Sierra Leone: Cross-Sectional Study"

Peter Bai James

Southern Cross University, Lismore, Australia

Related Articles:

Companion article: https://www.medrxiv.org/content/10.1101/2024.11.13.24317261v1

Companion article: https://med.jmirx.org/2025/1/e75127

Companion article: https://med.jmirx.org/2025/1/e68865

(JMIRx Med 2025;6:e75134) doi:10.2196/75134

KEYWORDS

academic bullying; junior doctors; Sierra Leone; mental health; professional development

This is a peer-review report for "Prevalence and Determinants of Academic Bullying Among Junior Doctors in Sierra Leone: Cross-Sectional Study."

Round 1 Review

Specific Comments

Major Comments

Introduction

I think the Introduction in this study [1] needs to be contextualized properly. Saying that bullying in the health care profession has not been looked at is largely correct, but the authors need to strengthen their argument by properly discussing the current literature on bullying in the Sierra Leone educational establishment and the limitations of the current literature as it relates to their topic of enquiry.

Please read the following:

- Osborne A, James PB, Bangura C, Tom Williams SM, Kangbai JB, Lebbieie, A. Bullying victimization among in-school adolescents in Sierra Leone: a cross-sectional analysis of the 2017 Sierra Leone Global School-Based Health Survey. *PLOS Glob Public Health*. Dec 22, 2023;3(12):e0002498. [doi: 10.1371/journal.pgph.0002498] [PMID: 38134001]
- Report on findings from school-related gender-based violence action research in schools and communities in Sierra Leone [2].

Methods

I wonder why the authors decided not to recruit all junior doctors who met their inclusion criteria, given that the list of junior doctors in the University of Sierra Leone Teaching Hospitals Complex at the time of data collection can be obtained from each of the constituent teaching hospitals. I know for a fact that the population of junior doctors is not so huge (less than 500). In other words, why did the authors just recruit all 160 junior doctors? Such data can be sourced from the Sierra Leone Medical and Dental Association or from the respective teaching hospital.

What informed the design of the questionnaire used? Why did the authors decide not to conduct any form of validation of the questionnaire (ie, externally or internally) to ensure it is appropriate for the context in which it is used?

This study was among junior doctors, but the authors mentioned registrars. A registrar is no longer a junior doctor. I may be wrong, but I strongly suggest that the authors provide a clear definition of what is the definition of junior doctor in Sierra Leone.

Discussion

I beg to disagree. A sample was calculated, and a probabilistic sampling method was used in this study, which means that it gives an equal chance for everyone to be chosen. Thus, the sample used is representative of junior doctors in the University of Sierra Leone Teaching Hospitals Complex. There are two ways to explain your finding: either the sample is not representative because the sampling was not probabilistic or the whole population should have been recruited, or the finding is correct (ie, there are no gender differences).

Minor Comments

The first two sentences of the third paragraph of the Introduction section: This has already been stated in the previous paragraph. This is just a repetition.

Conflicts of Interest

None declared.

References

1. Jalloh F, Bah AT, Kanu A, et al. Prevalence and determinants of academic bullying among junior doctors in Sierra Leone: cross-sectional study. JMIRx Med 2025;6:e68865. [doi: 10.2196/68865]
2. Report on findings from school-related gender-based violence action research in schools and communities in Sierra Leone. United Nations Girls' Education Initiative. URL: https://www.ungei.org/publication/report-findings-school-related-gender-based-violence-action-research-schools-and [accessed 2025-04-16]

Edited by S Tungjitviboonkun; submitted 28.03.25; this is a non–peer-reviewed article;accepted 28.03.25; published 22.05.25.

Please cite as:
James PB
Peer Review of "Prevalence and Determinants of Academic Bullying Among Junior Doctors in Sierra Leone: Cross-Sectional Study"
JMIRx Med 2025;6:e75134
URL: https://xmed.jmir.org/2025/1/e75134
doi:10.2196/75134

Peer Review of "Prevalence and Determinants of Academic Bullying Among Junior Doctors in Sierra Leone: Cross-Sectional Study"

Jenny Wilkinson

Metavision Institute, Brisbane City, Australia

Related Articles:

Companion article: https://www.medrxiv.org/content/10.1101/2024.11.13.24317261v1

Companion article: https://med.jmirx.org/2025/1/e75127

Companion article: https://med.jmirx.org/2025/1/e68865

(JMIRx Med 2025;6:e75135) doi:10.2196/75135

KEYWORDS

academic bullying; junior doctors; Sierra Leone; mental health; professional development

This is the peer-review report for "Prevalence and Determinants of Academic Bullying Among Junior Doctors in Sierra Leone: Cross-Sectional Study."

Round 2 Review

General Comments

This study [1] presents a survey of junior doctors in Sierra Leone hospitals and their experience of bullying and found high levels of bullying among the participants. Below are comments and suggestions for clarifying and strengthening the work.

Specific Comments

Major Comments

1. The author's definition of bullying and whether it was provided to participants is somewhat unclear. In the abstract, bullying is described as involving repeated behaviors, which aligns with the typical definition of bullying as an ongoing or repeated action. However, in the Methods section, participants were asked to respond based on any instance of various behaviors. While a single act of intimidation, for example, constitutes inappropriate behavior that should be addressed, it may not meet the standard definition of bullying. It is essential to clarify this distinction and ensure that participants also recognized the difference so that general poor behavior is not conflated with bullying.
2. Was sampling randomly, equally, or proportionally distributed across the four sites, and were there any analyses done based on site?
3. How was random sampling achieved?
4. Please comment on the reliability and validity of the instrument used to collect data. What literature was used to inform the development of the questions? Please include this information in the manuscript.
5. At the start of paragraph 3 of the Introduction, the authors refer to "other contexts"; it is unclear what contexts are being referred to in this and the preceding paragraph.
6. The Introduction and Discussion would be strengthened by more specific references to literature findings. I found the text in both a little superficial.
7. It is unclear whether the participants were reporting behaviors they personally experienced (ie, they were bullied) against behaviors they observed (ie, others being bullied).
8. Please provide clarification as to who is a "junior doctor." This journal has an international readership, and this term can be used differently in different countries, with "junior doctors" having different lengths of service. Please ensure this is clear within the body of the manuscript.
9. The description of the multiple regression seems a little excessive given the lack of statistical significance. This could be made more concise and simply refer readers to Table 3. Similarly, the authors should be cautious not to overemphasize these findings.
10. The list of references needs to be reviewed to ensure that all items have full bibliographic details.

Round 3 Review

The authors have addressed the review comment in their response, and this has been somewhat translated to the manuscript itself, noting that the lack of track changes, list of specific changes, or other highlights of manuscript revisions makes it difficult to see what changes were made. For example, while the comments regarding instrument development are addressed in the authors' response, it is unclear whether any changes have been made to the manuscript itself.

Conflicts of Interest

None declared.

Reference

1. Jalloh F, Bah AT, Kanu A, et al. Prevalence and determinants of academic bullying among junior doctors in Sierra Leone: cross-sectional study. JMIRx Med 2025;6:e68865. [doi: 10.2196/68865]

Edited by S Tungjitviboonkun; submitted 28.03.25; this is a non–peer-reviewed article;accepted 28.03.25; published 22.05.25.

Please cite as:
Wilkinson J
Peer Review of "Prevalence and Determinants of Academic Bullying Among Junior Doctors in Sierra Leone: Cross-Sectional Study"
JMIRx Med 2025;6:e75135
URL: https://xmed.jmir.org/2025/1/e75135
doi:10.2196/75135

Commentary on "Prevalence of Undiagnosed Hypertension Among Adult Displaced Individuals in Baidoa Camps, Somalia (Preprint)"

Anonymous

Related Articles:

Companion article: https://www.medrxiv.org/content/10.1101/2024.03.22.24304736v1

Companion article: https://med.jmirx.org/2025/1/e70265

(JMIRx Med 2025;6:e71041) doi:10.2196/71041

KEYWORDS

prevalence; undiagnosed; epidemiology; heart; cardiology; cardiovascular; cross-sectional; survey; questionnaires; hypertension; blood pressure; poverty; sedentary; displaced; refugee; Africa

This is a peer-review report submitted for the preprint "Prevalence of Undiagnosed Hypertension Among Adult Displaced Individuals in Baidoa Camps, Somalia." The preprint did not proceed to publication in JMIRx Med. In these cases, JMIRx-branded journals, acting as overlay journals for preprints, may publish peer-review reports as commentaries.

Round 1 Review

I commend the author for this study [1] on an important topic. However, here are a few comments to help improve the manuscript.

Title

1. The title needs some slight changes to improve clarity. For instance, what do you mean by "displaced individuals"? Would you rather state it as "internally displaced persons" or just "adults in Baidoa displacement camps"?
2. Use a uniform font for the title.

Introduction

1. Ensure a consistent referencing style throughout the manuscript.
2. In the sentence "Over the past few decades, the...," delete the bracket at the end of the statement.
3. Check the overall grammar of the text throughout the manuscript.
4. Regarding the burden of hypertension, provide more updated statistics on hypertension, using both global and regional data. Ensure a clear linkage and transition between the two because, as it stands right now, the statistics are scattered throughout the introduction, rendering it redundant.
5. Provide more context on the displaced populations and their specific vulnerabilities to hypertension to strengthen the rationale of the study. Discuss the factors therein.
6. The section would benefit from a discussion on the effects of hypertension.
7. Cite studies that have investigated hypertension among displaced populations, if any exist, or state the deficit if none.
8. Discuss any interventions and strategies that have been implemented to tackle the problem of hypertension in these communities and state the possible gaps before your objective.

Methods

1. Formatting issue: provide a heading for your Methods section.
2. As stated above, there is a need to improve the overall grammar.
3. Provide more detail regarding the inclusion criteria. For instance, was there a specific displacement duration that was considered (ie, the minimum amount of time spent in the camp so far)?
4. Provide a justification for the exclusion criteria.
5. Provide the reference for "The sample size for this study was determined...."
6. Add more detail regarding the validation of the questionnaire. Was it adopted from previous studies? Was it pretested?
7. Add detail on the measurement of blood pressure (BP). Who measured the BPs? Were they trained? How did you deal with white-coat hypertension? What was the interval between the different BP readings?

Results

1. Again, appropriate headings should be provided. Check the grammar.
2. Provide a more simplified and summarized Results section. For instance, "In this study, we enrolled 240 respondents, with a mean age...."
3. Table 1 is very confusing, especially the frequency and percentage columns. Clearly provide both the frequencies and percentages.
4. Add a key for Figure 2 to give better representation or just integrate the data represented into the text.

Discussion

1. Restate the objective at the start.
2. Provide a concise summary of key findings.
3. Thoroughly discuss the implications of the factors found to be significantly associated with hypertension.

Conflicts of Interest

None declared.

Reference

1. Jayte M. Prevalence of undiagnosed hypertension among adult displaced individuals in Baidoa camps, Somalia. medRxiv. Preprint posted online on Mar 26, 2024. [doi: 10.1101/2024.03.22.24304736]

Abbreviations

BP: blood pressure

Edited by E Meinert; submitted 08.01.25; this is a non–peer-reviewed article;accepted 08.01.25; published 03.06.25.

Please cite as:
Anonymous
Commentary on "Prevalence of Undiagnosed Hypertension Among Adult Displaced Individuals in Baidoa Camps, Somalia (Preprint)"
JMIRx Med 2025;6:e71041
URL: https://xmed.jmir.org/2025/1/e71041
doi:10.2196/71041

Peer Review of “Financial Feasibility of Developing Sustained-Release Incrementally Modified Drugs in Thailand’s Pharmaceutical Industry: Mixed Methods Study”

Parnnaphat Luksameesate, PhD

Chulalongkorn University, 254 Phaya Thai Rd, Bangkok, Thailand

Related Articles:

Companion article: https://www.medrxiv.org/content/10.1101/2024.07.29.24311184v1

Companion article: https://med.jmirx.org/2025/1/e77623

Companion article: https://med.jmirx.org/2025/1/e65978

(JMIRx Med 2025;6:e78090) doi:10.2196/78090

KEYWORDS

financial; economics; R&D; research and development; surveys; interviews; costs; revenue; policies; drugs; pharmaceuticals

This is the peer-review report for “Financial Feasibility of Developing Sustained-Release Incrementally Modified Drugs in Thailand’s Pharmaceutical Industry: Mixed Methods Feasibility Study.”

Round 1 Review

General Comments

This paper [1] provides valuable insights into how the Thai pharmaceutical industry should prepare for future developments. The results can be used as a reference to support decision-making and to guide the definition of regulations and processes in Thailand.

Specific Comments

Major Comments

1. Methods: Could you elaborate on how the 5 incrementally modified drug (IMD) experts were selected? Additionally, why was the number of experts limited to 5?

2. Tables 1 and 2: Please replace the term “Literature Review” with the specific author names and the corresponding year (Anno Domini).

3. Table 3: The values of US $1.46 million and US $18.6 million refer to the research and development costs only, correct? These values do not reflect the total cost of developing IMDs (refer to Table 2).

4. Since most of the numbers come from expert input, how do you ensure that these numbers are valid and accurately reflect real-world situations? It may be helpful to provide more information about the characteristics and qualifications of the key informants to support their credibility.

Minor Comments

5. Please ensure that all abbreviations are defined the first time they appear in the document. For example, “IMD” should be written out as “Innovative Medical Devices (IMD)” when it is first mentioned, particularly in the introduction.

Conflicts of Interest

None declared.

Reference

1. Laichapis M, Sakulbumrungsil R, Udomaksorn K, et al. Financial feasibility of developing sustained-release incrementally modified drugs in Thailand’s pharmaceutical industry: mixed methods feasibility study. JMIRx Med 2025;6:e65978. [doi: 10.2196/65978]

Abbreviations

IMD: incrementally modified drug

Edited by A Grover; submitted 26.05.25; this is a non–peer-reviewed article;accepted 26.05.25; published 01.07.25.

Please cite as:
Luksameesate P
Peer Review of "Financial Feasibility of Developing Sustained-Release Incrementally Modified Drugs in Thailand's Pharmaceutical Industry: Mixed Methods Study"
JMIRx Med 2025;6:e78090
URL: https://xmed.jmir.org/2025/1/e78090
doi:10.2196/78090

Peer Review of “Financial Feasibility of Developing Sustained-Release Incrementally Modified Drugs in Thailand’s Pharmaceutical Industry: Mixed Methods Study”

Elena Shkarupeta

Voronezh State Technical University, Moskovskiy Prospekt, 14, Voronezh, Russian Federation

Related Articles:

Companion article: https://www.medrxiv.org/content/10.1101/2024.07.29.24311184v1

Companion article: https://med.jmirx.org/2025/1/e77623

Companion article: https://med.jmirx.org/2025/1/e65978

(JMIRx Med 2025;6:e77627) doi:10.2196/77627

KEYWORDS

financial; economics; R&D; research and development; surveys; interviews; costs; revenue; policies; drugs; pharmaceuticals

This is the peer-review report for “Financial Feasibility of Developing Sustained-Release Incrementally Modified Drugs in Thailand’s Pharmaceutical Industry: Mixed Methods Feasibility Study.”

Round 1 Review

General Comments

This paper [1] presents a thorough analysis of the financial feasibility of developing incrementally modified drugs (IMDs) within the Thai pharmaceutical industry. It aligns well with Thailand’s National Strategic Master Plan and provides valuable insights for stakeholders regarding investment decisions and policy development. The mixed methods approach, including financial modeling, surveys, and interviews, lends credibility to the findings, while the focus on sustained-release dosage forms highlights a specific and practical application. The paper is well structured and contributes meaningfully to the discussion on enhancing local pharmaceutical capabilities. However, there are areas where clarity, presentation, and depth can be improved to strengthen its impact.

Specific Comments

Major Comments

1. Clarity in objectives: While the paper provides an extensive background on Thailand’s pharmaceutical landscape, the research objectives could be more explicitly stated at the beginning of the introduction to guide the reader more effectively.

2. Discussion of results: The discussion section could delve deeper into comparing the financial feasibility of IMDs with other pharmaceutical products, especially generic drugs, to highlight the broader implications of the findings.

3. Policy recommendations: Although the paper suggests policy recommendations, it would benefit from providing concrete examples of how these policies have been successfully implemented in other regions or industries. This would add depth and context to the recommendations.

4. References and citation quality: The paper relies on only 15 references, which is insufficient for a study of this scope. Furthermore, only a few of these references are from peer-reviewed scientific journals, while the rest are reports and secondary sources. This significantly weakens the academic foundation of the study. It is strongly recommended to update the references section by incorporating recent, high-quality, and peer-reviewed articles.

Minor Comments

5. Terminology consistency: Terms like “incrementally modified drugs” and “IMDs” should be consistently used throughout the text to avoid confusion.

6. Figures and tables: Ensure all figures and tables are adequately labeled and referenced in the text. For instance, the presentation of financial data could be enhanced with clearer visualizations.

7. Formatting and grammar: Minor grammatical errors and formatting inconsistencies (eg, use of citations and spacing) should be addressed for a polished presentation.

8. Abstract refinement: The abstract could be more concise, emphasizing key findings and policy implications without overly detailed descriptions of methods.

9. Future research directions: Including a section on future research directions would enhance the paper’s utility for academics and policy makers.

Reference

1. Laichapis M, Sakulbumrungsil R, Udomaksorn K, et al. Financial feasibility of developing sustained-release incrementally modified drugs in Thailand's pharmaceutical industry: mixed methods feasibility study. JMIRx Med 2025;6:e65978. [doi: 10.2196/65978]

Abbreviations

IMD: incrementally modified drug

Edited by A Grover; submitted 16.05.25; this is a non–peer-reviewed article;accepted 16.05.25; published 01.07.25.

Please cite as:
Shkarupeta E
Peer Review of "Financial Feasibility of Developing Sustained-Release Incrementally Modified Drugs in Thailand's Pharmaceutical Industry: Mixed Methods Study"
JMIRx Med 2025;6:e77627
URL: https://xmed.jmir.org/2025/1/e77627
doi:10.2196/77627

Peer Review of "Improving Tuberculosis Detection in Chest X-Ray Images Through Transfer Learning and Deep Learning: Comparative Study of Convolutional Neural Network Architectures"

Natthapong Nanthasamroeng

Ubon Ratchathani Rajabhat University, Ubon Ratchathani, Thailand

Related Articles:

Companion article: https://www.medrxiv.org/content/10.1101/2024.08.02.24311396v1

Companion article: https://med.jmirx.org/2025/1/e 77221

Companion article: https://med.jmirx.org/2025/1/e66029

(JMIRx Med 2025;6:e77174) doi:10.2196/77174

KEYWORDS

tuberculosis detection; tuberculosis; TB; chest x-ray classification; diagnostic imaging; radiology; medical imaging; convolutional neural networks; data augmentation; deep learning; early warning; early detection; comparative study

This is a peer-review report for "Improving Tuberculosis Detection in Chest X-Ray Images Through Transfer Learning and Deep Learning: Comparative Study of Convolutional Neural Network Architectures."

Round 1 Review

General Comments

The manuscript [1] presents a study that evaluates the performance of various convolutional neural network architectures—namely, VGG16, VGG19, ResNet50, ResNet101, ResNet152, and Inception-ResNet-V2—in classifying chest x-ray images to detect tuberculosis (TB). The authors compare the models' classification accuracy, precision, recall, F_1-score, and area under the receiver operating characteristic curve, concluding that VGG16 outperforms the others with high accuracy and efficiency. They also assess the impact of data augmentation, finding it does not improve model performance due to sufficient diversity in the original dataset.

Specific Comments

1. The dataset includes a large imbalance between TB-positive and TB-negative images (700 vs 3500). Explain how this imbalance was addressed beyond augmentation or whether balancing techniques like oversampling were considered.
2. While each architecture's parameters are listed, there is no in-depth discussion on why these specific parameters (eg, dropout rates, learning rates) were selected.
3. The conclusion that data augmentation did not improve performance lacks specific references to possible reasons.
4. While computational time for each model is reported, further analysis of the practical implications, such as cost-effectiveness for clinical settings, is missing.
5. The manuscript mentions transfer learning with pretrained ImageNet weights, but there is limited information on why this was the chosen approach versus training from scratch.
6. Throughout the Results section, adding comparative charts or visual aids for each model's performance across metrics like accuracy, precision, and area under the receiver operating characteristic curve would improve readability.
7. The Conclusion could benefit from a clearer statement on how these findings advance the field of TB detection in medical imaging.

Conflicts of Interest

None declared.

Reference

1. Mirugwe A, Tamale L, Nyirenda J. Improving Tuberculosis Detection in Chest X-Ray Images Through Transfer Learning and Deep Learning: Comparative Study of Convolutional Neural Network Architectures. JMIRx Med 2025;6:e66029. [doi: 10.2196/66029]

Abbreviations

TB: tuberculosis

Edited by S Amal; submitted 08.05.25; this is a non–peer-reviewed article;accepted 08.05.25; published 01.07.25.

Please cite as:
Nanthasamroeng N
Peer Review of "Improving Tuberculosis Detection in Chest X-Ray Images Through Transfer Learning and Deep Learning: Comparative Study of Convolutional Neural Network Architectures"
JMIRx Med 2025;6:e77174
URL: https://xmed.jmir.org/2025/1/e77174
doi:10.2196/77174

Peer Review of "Improving Tuberculosis Detection in Chest X-Ray Images Through Transfer Learning and Deep Learning: Comparative Study of Convolutional Neural Network Architectures"

Rapeepan Pitakaso
University of Vienna, remove, Vienna, Austria

Related Articles:

Companion article: https://www.medrxiv.org/content/10.1101/2024.08.02.24311396v1

Companion article: https://med.jmirx.org/2025/1/e 77221

Companion article: https://med.jmirx.org/2025/1/e66029

(JMIRx Med 2025;6:e77171) doi:10.2196/77171

KEYWORDS

tuberculosis detection; tuberculosis; TB; chest x-ray classification; diagnostic imaging; radiology; medical imaging; convolutional neural networks; data augmentation; deep learning; early warning; early detection; comparative study

This is a peer-review report for "Improving Tuberculosis Detection in Chest X-Ray Images Through Transfer Learning and Deep Learning: Comparative Study of Convolutional Neural Network Architectures."

Round 1 Review

General Comments

Clarity and Structure

The paper [1] presents a comprehensive overview of the methods and results but can benefit from clearer transitions between sections. For instance, adding brief connecting sentences at the end of each section would help guide the reader into the next topic.

Consider reorganizing the "Discussion" section to first summarize the key findings before delving into their implications. This will reinforce the reader's understanding of the main outcomes.

Writing Style

Aim for more active voice usage to enhance readability. For example, change "It was observed that VGG16 outperformed other models" to "We observed that VGG16 outperformed other models."

Simplify overly technical or long sentences to improve readability. Breaking complex sentences into two simpler ones can make the content easier to follow.

Specific Comments by Section

Abstract

Sentence clarification: The phrase "necessitating more efficient and accurate diagnostic methods" could be expanded to briefly indicate why current methods are insufficient.

Results detail: When mentioning model performance, briefly state why VGG16's superior performance is significant compared to others.

Introduction

Background information: The explanation of the global tuberculosis burden is informative, but it could benefit from briefly mentioning current limitations in artificial intelligence–based tuberculosis detection in developing countries.

Motivation clarification: Ensure that the motivation for choosing specific convolutional neural network architectures is clearly linked to gaps in existing literature.

Methods

Preprocessing details: The detailed explanation of normalization and data augmentation is excellent, but it might be beneficial to briefly mention how these choices align with previous research findings or unique aspects of this study.

Transfer learning: Include a brief comparison of why transfer learning was chosen over training models from scratch.

Results

Visualization: The table summarizing model performance is comprehensive, but consider including a concise narrative to describe key trends observed in the data.

Analysis clarification: When discussing why data augmentation did not enhance performance, elaborate on how this aligns with or contradicts findings from other studies.

Discussion

Comparison with previous studies: Add a few sentences comparing the results with existing studies that used the same models or datasets to provide context.

Implications: Discuss the practical implications of using VGG16 in resource-constrained environments where computational efficiency is crucial.

Conclusion

Highlight novelty: Emphasize what makes this study's approach unique, such as the use of specific architectures on a larger dataset, and how this adds to the current body of knowledge.

Future work suggestions: Include more detailed recommendations for future studies, potentially suggesting how to further leverage data augmentation strategies.

Grammar and Language

Sentence revisions: original: "It is observed that the VGG16 consistently performed better than other models." Revised: "We observed that VGG16 consistently performed better than the other models."

Punctuation: Ensure commas are consistently used after introductory phrases (eg, "In this study, we propose...").

Word choice: Replace terms like "aimed to assess" with "assessed" to make sentences more concise.

Technical Aspects

Hyperparameter details: Include a brief rationale for choosing the specific hyperparameters in Table 1 to enhance the reader's understanding.

Training environment: Specify why the computational setup (eg, graphics processing unit details) was chosen and how it impacted training efficiency.

Final Suggestions

Proofreading: Ensure that each section is proofread for minor grammatical errors or inconsistencies.

Figures and tables: Verify that all figures and tables have descriptive captions, and refer to them within the text to maintain flow.

Conflicts of Interest

None declared.

Reference

1. Mirugwe A, Tamale L, Nyirenda J. Improving Tuberculosis Detection in Chest X-Ray Images Through Transfer Learning and Deep Learning: Comparative Study of Convolutional Neural Network Architectures. JMIRx Med 2025;6:e66029. [doi: 10.2196/66029]

Edited by S Amal; submitted 08.05.25; this is a non–peer-reviewed article;accepted 08.05.25; published 01.07.25.

Please cite as:
Pitakaso R
Peer Review of "Improving Tuberculosis Detection in Chest X-Ray Images Through Transfer Learning and Deep Learning: Comparative Study of Convolutional Neural Network Architectures"
JMIRx Med 2025;6:e77171
URL: https://xmed.jmir.org/2025/1/e77171
doi:10.2196/77171

Peer Review of “Advancing Early Detection of Major Depressive Disorder Using Multisite Functional Magnetic Resonance Imaging Data: Comparative Analysis of AI Models”

Anonymous

Related Articles:

Companion article: https://www.medrxiv.org/content/10.1101/2024.08.13.24311933v1

Companion article: https://med.jmirx.org/2025/1/e75617

Companion article: https://med.jmirx.org/2025/1/e65417

(JMIRx Med 2025;6:e76744) doi:10.2196/76744

KEYWORDS

major depressive disorder; machine learning; functional MRI; early detection; artificial intelligence; psychiatry

This is a peer-review report for “Advancing Early Detection of Major Depressive Disorder Using Multisite Functional Magnetic Resonance Imaging Data: Comparative Analysis of AI Models.”

Round 1 Review

This paper [1] addresses a relevant and important topic in psychiatric research. The authors aim to develop and compare machine learning models for early detection of major depressive disorder using functional magnetic resonance imaging (fMRI) data, which is a novel and promising approach. The study appears to be well structured and utilizes an appropriate set of methodologies to evaluate the machine learning models. However, some issues need to be addressed before the manuscript can be considered for publication.

Specific Comments

Major Comments

- Interpretability of artificial intelligence (AI) models: While the paper discusses the models’ performance, it would benefit from further elaboration on the interpretability of the models, particularly the clinical relevance of Shapley additive explanations values and activation maximization findings. Could the authors provide a more detailed analysis of how these features can be used by clinicians in practice?
- Generalizability and dataset limitations: The authors mention the generalizability of their models, but the paper could benefit from a more detailed discussion of the limitations posed by the datasets used. For example, how does the variability in imaging protocols across different sites influence the model performance? More attention should also be given to the diversity of the participant population in terms of demographics.
- Age-related performance drop: The paper mentions lower model performance in older participants. This is a significant finding and should be explored further. Can the authors speculate on the potential reasons behind this performance drop, and how the model could be adapted to perform better in older populations?

Minor Comments

- Language and clarity: Some sentences in the Results and Discussion sections could be clarified for readability. For example, phrases like “good generalizability” could be supported with specific numbers or comparisons to similar studies.
- Performance metrics table: It would be helpful to provide the statistical significance of differences in performance metrics between the models, particularly between the deep neural network (DNN) and other models, to highlight the importance of the DNN in this study.
- Ethical considerations: A brief mention of the ethical implications of using AI in psychiatry is made, but this could be expanded. Ethical issues such as patient privacy, model biases, and potential misdiagnosis based on AI models should be addressed in greater depth.

Round 2 Review

The paper presents an analysis of several AI models (support vector machine, random forest, gradient boosting machine, and DNN) for the early detection of major depression disorder using multisite fMRI data. The study offers valuable insights into both predictive performance and model interpretability. It is commendable that the authors leverage a diverse dataset and employ robust validation techniques (eg, 5-fold cross-validation and external validation) to assess model generalizability. However, there are areas—particularly in methodological clarity and discussion of clinical translation—that would benefit from further refinement.

Major Comments

Methodological Details and Preprocessing

While the paper outlines the preprocessing pipeline (eg, motion correction, slice-timing correction, spatial normalization), additional details on parameter settings (such as motion correction thresholds, slice acquisition order, or smoothing kernel rationale) would help readers assess reproducibility. Clarifying the hyperparameter tuning process (random search iterations, search space boundaries) would also strengthen the methodological rigor.

Data Heterogeneity and Generalizability

The study uses fMRI data from three public datasets, which is a strength in terms of diversity. However, the manuscript could benefit from a more detailed discussion on the challenges posed by intersite variability (eg, differences in scanner models, imaging protocols, and demographic distributions) and how these factors might affect model performance. Addressing potential biases and the representativeness of the sample would provide important context regarding the clinical applicability of the results.

Interpretability and Clinical Integration

The inclusion of feature importance and Shapley additive explanations analyses is a positive step toward interpretability. Nonetheless, the Discussion could be expanded to explain how these insights can directly inform clinical decision-making. For example, a deeper exploration of how the identified neural connectivity patterns relate to established neurobiological theories of major depressive disorder—and what this means for potential treatment interventions—would enhance the translational impact of the work.

Minor Comments

Clarity and Language

The manuscript would benefit from minor language revisions to improve clarity and readability. Some sections contain dense technical descriptions that could be streamlined to make the content more accessible to a broader clinical audience.

Figures and Tables

Ensure that all figures (especially the model performance comparison chart) and tables are clearly labeled and of sufficient resolution. Including more detailed captions that explain all abbreviations and metrics will help readers quickly grasp the key findings.

Discussion Section

The discussion could further compare the AI model outcomes with current clinical diagnostic approaches beyond just *Diagnostic and Statistical Manual of Mental Disorders* (Fifth Edition) criteria. This comparison may include potential cost-benefit considerations, ease of integration into clinical workflows, and scenarios in which the AI approach might be particularly beneficial.

Future Directions

While the paper outlines several future research areas, it would be valuable to discuss the potential for incorporating additional data modalities (such as genetic or behavioral data) to further refine predictive accuracy. Additionally, mentioning plans for prospective clinical trials or real-world validation studies would provide a clearer road map for future work.

Conflicts of Interest

None declared.

Reference

1. Mansoor M, Ansari K. Advancing Early Detection of Major Depressive Disorder Using Multisite Functional Magnetic Resonance Imaging Data: Comparative Analysis of AI Models. JMIRx Med 2025;6:e65417. [doi: 10.2196/65417]

Abbreviations

AI: artificial intelligence
DNN: deep neural network
fMRI: functional magnetic resonance imaging

Edited by CN Hang; submitted 29.04.25; this is a non–peer-reviewed article;accepted 29.04.25; published 15.07.25.

Please cite as:
Anonymous
Peer Review of "Advancing Early Detection of Major Depressive Disorder Using Multisite Functional Magnetic Resonance Imaging Data: Comparative Analysis of AI Models"
JMIRx Med 2025;6:e76744
URL: https://xmed.jmir.org/2025/1/e76744
doi:10.2196/76744

Peer Review of “Advancing Early Detection of Major Depressive Disorder Using Multisite Functional Magnetic Resonance Imaging Data: Comparative Analysis of AI Models”

Anonymous

Related Articles:

Companion article: https://www.medrxiv.org/content/10.1101/2024.08.13.24311933v1

Companion article: https://med.jmirx.org/2025/1/e75617

Companion article: https://med.jmirx.org/2025/1/e65417

(JMIRx Med 2025;6:e76746) doi:10.2196/76746

KEYWORDS

major depressive disorder; machine learning; functional MRI; early detection; artificial intelligence; psychiatry

This is a peer-review report for “Advancing Early Detection of Major Depressive Disorder Using Multisite Functional Magnetic Resonance Imaging Data: Comparative Analysis of AI Models.”

Round 1 Review

General Comments

The paper [1] “Advancing Early Detection of Major Depressive Disorder: A Comparative Analysis of AI Models Using Multi-Site Functional MRI Data” examines a very relevant mental health disorder. The purpose of this paper is to identify the best artificial intelligence (AI) model for predicting early detection using a more comprehensive and versatile dataset. The paper’s contribution to psychiatry could be to provide the best AI model with specific features that can be generalized to a larger population. The paper also included a comparison of health control measures, which could improve the prediction’s accuracy. The manuscript’s most notable feature is the inclusion of 2-year longitudinal data for the early detection of major depressive disorder (MDD).

Major Comments

1. The manuscript’s goal is to provide early but accurate detection of MDD to help with diagnosis. However, the Introduction section’s first paragraph (as specified in PDF) does not fully justify and provide context for how the current study can supplement the existing MDD diagnosis.
2. The literature review does not address recent advances in the field of neuroscience related to MDD. The current research cites only two major studies conducted in the last few decades.
3. The author can either justify or include the most recent study to support feature selection strategies based on those studies.
4. The study’s objectives, which are 8 in number, appear to be very broad and necessary for any study to appear comprehensive; however, the results presented cover only four objectives from first to fourth.
5. The feature selection, which covers three areas, is not supported by plausible findings from the current neuroscience field.
6. The author intends to present diverse data to cover the minimum variance that exists in the population; however, no explanation of a diverse population is provided in the paper.
7. The literature review presented in the manuscript could be more rigorous, first explaining the gaps in the current literature regarding the use of machine learning and deep neural networks in the detection of MDD, then explaining the best feature and detection method for MDD, and finally explaining the findings.
8. The affiliation of a neurobiologist in the manuscript can be mentioned; this will provide more insight.
9. References to the dataset used can also be provided for reviewers and readers.

Conflicts of Interest

None declared.

Reference

1. Mansoor M, Ansari K. Advancing Early Detection of Major Depressive Disorder Using Multisite Functional Magnetic Resonance Imaging Data: Comparative Analysis of AI Models. JMIRx Med 2025;6:e65417. [doi: 10.2196/65417]

Abbreviations

AI: artificial intelligence
MDD: major depressive disorder

Edited by CN Hang; submitted 29.04.25; this is a non–peer-reviewed article;accepted 29.04.25; published 15.07.25.

Please cite as:
Anonymous
Peer Review of "Advancing Early Detection of Major Depressive Disorder Using Multisite Functional Magnetic Resonance Imaging Data: Comparative Analysis of AI Models"
JMIRx Med 2025;6:e76746
URL: https://xmed.jmir.org/2025/1/e76746
doi:10.2196/76746

Peer Review of “Advancing Early Detection of Major Depressive Disorder Using Multisite Functional Magnetic Resonance Imaging Data: Comparative Analysis of AI Models”

Anonymous

Related Articles:

Companion article: https://www.medrxiv.org/content/10.1101/2024.08.13.24311933v1

Companion article: https://med.jmirx.org/2025/1/e75617

Companion article: https://med.jmirx.org/2025/1/e65417

(JMIRx Med 2025;6:e76747) doi:10.2196/76747

KEYWORDS

major depressive disorder; machine learning; functional MRI; early detection; artificial intelligence; psychiatry

This is a peer-review report for “Advancing Early Detection of Major Depressive Disorder Using Multisite Functional Magnetic Resonance Imaging Data: Comparative Analysis of AI Models.”

Round 1 Review

Specific Comments

Major Comments

1. This paper [1] provides sufficient information about major depressive disorder and the potential of artificial intelligence (AI); it could benefit from a more detailed comparison with the existing literature. How does the present study build on or extend previous work? Additional details on why previous AI studies have not focused on early detection could help contextualize the research gap you are addressing.

Minor Comments

2. It’s also important to emphasize that AI should complement, rather than replace, clinical expertise.

Conflicts of Interest

None declared.

Reference

1. Mansoor M, Ansari K. Advancing Early Detection of Major Depressive Disorder Using Multisite Functional Magnetic Resonance Imaging Data: Comparative Analysis of AI Models. JMIRx Med 2025;6:e65417. [doi: 10.2196/65417]

Abbreviations

AI: artificial intelligence

Edited by CN Hang; submitted 29.04.25; this is a non–peer-reviewed article;accepted 29.04.25; published 15.07.25.

Please cite as:
Anonymous
Peer Review of “Advancing Early Detection of Major Depressive Disorder Using Multisite Functional Magnetic Resonance Imaging Data: Comparative Analysis of AI Models”
JMIRx Med 2025;6:e76747
URL: https://xmed.jmir.org/2025/1/e76747
doi:10.2196/76747

Peer Review of "Monte Carlo Dose Estimation of Absorbed Dose to the Hematopoietic Stem Cell Layer of the Bone Marrow Assuming Nonuniform Distribution Around the Vascular Endothelium of the Bone Marrow: Simulation and Analysis Study"

Randa Salah Gomaa Mahmoud

Zagazig University, Zagazig, Egypt

Related Articles:

Companion article: https://www.medrxiv.org/content/10.1101/2024.10.19.24315800v1

Companion article: https://med.jmirx.org/2025/1/e77812

Companion article: https://med.jmirx.org/2025/1/e68029

(JMIRx Med 2025;6:e77775) doi:10.2196/77775

KEYWORDS

stem cells; radiation; bone marrow; nuclides; noble gases

This is the peer-review report for "Monte Carlo Dose Estimation of Absorbed Dose to the Hematopoietic Stem Cell Layer of the Bone Marrow Assuming Nonuniform Distribution Around the Vascular Endothelium of the Bone Marrow: Simulation and Analysis Study."

Round 1 Review

General Comments

In this study [1], a geometric model of trabecular bone and bone marrow tissue was constructed at the micrometer scale, assuming that the hematopoietic stem cells layer was localized in the perivascular hematopoietic stem cell layer of the sinusoids. The absorbed doses of the stem cell layer from blood and trabecular bone sources were then estimated for selected β nuclides, α nuclides, and noble gases and compared with the specific absorbed fractions (SAFs) values of International Commission on Radiological Protection (ICRP) 60 and 103. It was concluded that the absorbed doses from the bone marrow and blood sources were greater than those from trabecular bone sources for α nuclides, and the total absorbed dose was lower than that estimated from the current ICRP models.

Specific Comments

1. The results were tabulated; however, it was not clear how the comparison between the Particle and Heavy Ion Transport System, ICRP 60, and ICRP 103 was performed, what test was used, and the level of significance. Even in Table 7 that summarizes the results, this is not clear.
2. The abbreviations throughout the article need to be identified. It is recommended to add an abbreviation section to the article.
3. The abstract section is better structured as Background, Objectives, Methods, Results, and Conclusion.
4. In the abstract section, the authors mentioned that the absorbed doses to the bone marrow obtained from the model calculations were not significantly different from ICRP 60 and ICRP 103 for β nuclides. Still, they were much lower than previously estimated for α nuclides. Going through the study, it was not clear how this significant difference was assessed. Please revise and clarify.
5. The abbreviation "SAFs" in the keyword section and the last paragraph of the Introduction section should be identified as the "specific absorbed fractions."
6. The abbreviation "PHITS" in the keyword section and the first line of the fourth page should be identified as "Particle and Heavy Ion Transport System."
7. The abbreviation "keV" in the last line of the second paragraph of the seventh page should be identified as "kilo electron-volt."
8. In the last line of the second paragraph of the seventh page, please identify "Bremsstrahlung" as a type of X-radiation emitted by charged particles when they collide or are near an atomic nucleus.
9. The abbreviation "EGS" in the last line of the second paragraph of the seventh page should be identified as "Electron Gamma Shower."
10. The abbreviation "Bq" in the first line of the last paragraph of the seventh page should be identified as "The International System of Units (SI) unit of radionuclide activity is the becquerel (Bq); 1 Bq = 1 transformation/second."
11. First line, page 10: Please correct "131" to "131I."
12. Page 16, Discussion section, last line of the first paragraph: The authors mentioned that the number of decays in each compartment changed significantly; how did the authors

assess this significant change and conclude it? Please explain the tests used for comparison.

13. Page 16, Discussion section, eighth line of the second paragraph: Please revise “ICRP133 SAF” (mentioned in the Results section as “ICRP103 SAF”).
14. Page 17, last line of the first paragraph: “Sakota et al” should be corrected to “Sakoda et al.”

Round 2 Review

General Comments

All the comments were professionally addressed.

Conflicts of Interest

None declared.

Reference

1. Kobayashi N. Monte Carlo dose estimation of absorbed dose to the hematopoietic stem cell layer of the bone marrow assuming nonuniform distribution around the vascular endothelium of the bone marrow: simulation and analysis study. JMIRx Med 2025;6:e68029. [doi: 10.2196/68029]

Abbreviations

ICRP: International Commission on Radiological Protection
SAF: specific absorbed fraction
SI: International System of Units

Edited by A Grover; submitted 19.05.25; this is a non–peer-reviewed article;accepted 19.05.25; published 16.07.25.

Please cite as:
Mahmoud RSG
Peer Review of “Monte Carlo Dose Estimation of Absorbed Dose to the Hematopoietic Stem Cell Layer of the Bone Marrow Assuming Nonuniform Distribution Around the Vascular Endothelium of the Bone Marrow: Simulation and Analysis Study”
JMIRx Med 2025;6:e77775
URL: https://xmed.jmir.org/2025/1/e77775
doi:10.2196/77775

Peer Review of "Monte Carlo Dose Estimation of Absorbed Dose to the Hematopoietic Stem Cell Layer of the Bone Marrow Assuming Nonuniform Distribution Around the Vascular Endothelium of the Bone Marrow: Simulation and Analysis Study"

Maha Gasmi

Manouba University, Manouba, Tunisia

Related Articles:

Companion article: https://www.medrxiv.org/content/10.1101/2024.10.19.24315800v1

Companion article: https://med.jmirx.org/2025/1/e77812

Companion article: https://med.jmirx.org/2025/1/e68029

(JMIRx Med 2025;6:e77776) doi:10.2196/77776

KEYWORDS

stem cells; radiation; bone marrow; nuclides; noble gases

This is the peer-review report for "Monte Carlo Dose Estimation of Absorbed Dose to the Hematopoietic Stem Cell Layer of the Bone Marrow Assuming Nonuniform Distribution Around the Vascular Endothelium of the Bone Marrow: Simulation and Analysis Study."

Round 1 Review

Abstract Section

1. The manuscript's [1] abstract begins with a statement about hematopoietic stem cells' proximity to sinusoidal capillaries but does not clarify why this spatial distribution is relevant for radiation dosimetry until later in the text. A clearer explanation linking the hematopoietic stem cell location with the dosimetric model limitations would better engage readers unfamiliar with the topic.

2. Some sentences are overly complex, especially in the Introduction and Conclusion. Simplifying the language or splitting ideas across multiple sentences could improve readability.

3. The abstract lacks methodological detail regarding how the model calculations were performed. Including brief specifics about the model's approach, particularly the role of computed tomography imaging if applicable, would improve transparency and give context to the reported findings.

4. The results comparing the absorbed doses for α and β nuclides are presented with limited interpretation. The abstract states that doses for β nuclides were similar to International Commission on Radiological Protection estimates, while those for α nuclides were much lower, yet there is no explanation for the potential reasons behind these differences. Offering a brief discussion or hypothesis, even speculative, would enrich the reader's understanding.

Introduction Section

5. The Introduction could benefit from a clearer structure. Currently, it presents information about various models and dosimetric approaches in a somewhat fragmented manner.

6. Certain technical terms such as "surrogate target," "trabecular bone surface," "endosteum," and "standard absorbed fraction" may benefit from concise explanations or definitions. For instance, briefly defining "surrogate target" would help those unfamiliar with dosimetry or radiobiology terminology.

Method Section

7. The study uses an intricate geometric model based on JM-103 data, Particle and Heavy Ion Transport System software, and Japan Atomic Energy Agency guidelines to simulate the cervical vertebrae trabecular bone. This choice is reasonable given the need for anatomical detail in dosimetry but may limit generalizability since the cervical vertebrae structure might not fully represent other bone marrow sites.

The description could benefit from clarifying why the JM-103 model was chosen over other models or datasets, particularly those that could include bone tissues beyond the cervical vertebrae.

Discussion Section

8. Despite noting the need for micro–computed tomgraphy–based models, the authors do not discuss how current limitations might impact dose estimation accuracy, especially for complex geometries in the trabecular bone. A clearer explanation of how simplified geometric assumptions may

influence absorbed dose calculations would provide a balanced view of the model's limitations.

Conflicts of Interest

None declared.

Reference

1. Kobayashi N. Monte Carlo dose estimation of absorbed dose to the hematopoietic stem cell layer of the bone marrow assuming nonuniform distribution around the vascular endothelium of the bone marrow: simulation and analysis study. JMIRx Med 2025;6:e68029. [doi: 10.2196/68029]

Edited by A Grover; submitted 19.05.25; this is a non–peer-reviewed article;accepted 19.05.25; published 16.07.25.

Please cite as:
Gasmi M
Peer Review of "Monte Carlo Dose Estimation of Absorbed Dose to the Hematopoietic Stem Cell Layer of the Bone Marrow Assuming Nonuniform Distribution Around the Vascular Endothelium of the Bone Marrow: Simulation and Analysis Study"
JMIRx Med 2025;6:e77776
URL: https://xmed.jmir.org/2025/1/e77776
doi:10.2196/77776

Peer Review of "Assessment of SARC-F Sensitivity for Probable Sarcopenia Among Community-Dwelling Older Adults: Cross-Sectional Questionnaire Study"

Anonymous

Related Articles:

Companion article: https://www.medrxiv.org/content/10.1101/2023.10.31.23297840v2

Companion article: https://med.jmirx.org/2025/1/e77497

Companion article: https://med.jmirx.org/2025/1/e54475

(JMIRx Med 2025;6:e78552) doi:10.2196/78552

KEYWORDS

sarcopenia; neuromuscular; screening; community; scale; measure; questionnaires; diagnosis; gerontology; older adults; muscular

This is the peer-review report for "Assessment of SARC-F Sensitivity for Probable Sarcopenia Among Community-Dwelling Older Adults: Cross-Sectional Questionnaire Study."

Round 1 Review

General Comments

The authors [1] present an intriguing and clinically valuable finding through their receiver operating characteristic (ROC) curve analysis, suggesting that a SARC-F (strength, assistance with walking, rising from a chair, climbing stairs, and falls) score of ≥2 may serve as a new cutoff value for screening probable sarcopenia. This conclusion has significant potential for improving clinical practice by enhancing early detection.

However, the study is based on a relatively small sample size of 204 community-dwelling older adults, and it is unclear if the data were collected from a single center. This limitation raises concerns about the generalizability of the findings to a broader population. I believe the authors could strengthen their argument by conducting additional analyses to address these limitations and provide more robust evidence.

Major Comments

1. Introduction: Add a discussion on current research gaps (eg, sarcopenia screening) and clearly explain how your study addresses these gaps.
2. Methods: Include additional clinical outcomes such as muscle function, sarcopenia-related symptoms, or quality of life, and compare how thresholds of ≥2 and ≥4 perform in relation to these outcomes.
3. Results: Provide more detailed basic characteristics of participants and compare these between thresholds of ≥2 and ≥4, referring to Malmstrom et al [2] for guidance.
4. Discussion: Update the Discussion to integrate insights from the new results, focusing on the implications of the revised threshold for clinical practice and your limitations.

Round 2 Review

Thank you for your revisions. I understand that due to the lack of relevant data, you were unable to expand your data analysis. I am pleased to see the addition of Tables 3 and 4 for the subgroup analysis; however, these two tables could be combined. Additionally, you may consider placing the ROC curves from Figures 1 and 2 into a single figure. Using software like MedCalc or SPSS to compare the areas under the different ROC curves would add more depth to the Results section.

Conflicts of Interest

None declared.

References

1. Propst D, Biscardi L, Dornemann T. Assessment of SARC-F sensitivity for probable sarcopenia among community-dwelling older adults: cross-sectional questionnaire study. JMIRx Med 2025;6:e54475. [doi: 10.2196/54475]
2. Malmstrom TK, Miller DK, Simonsick EM, Ferrucci L, Morley JE. SARC-F: a symptom score to predict persons with sarcopenia at risk for poor functional outcomes. J Cachexia Sarcopenia Muscle 2016 Mar;7(1):28-36. [doi: 10.1002/jcsm.12048] [Medline: 27066316]

Abbreviations

ROC: receiver operating characteristic
SARC-F: strength, assistance with walking, rising from a chair, climbing stairs, and falls

Edited by A Schwartz; submitted 04.06.25; this is a non–peer-reviewed article;accepted 04.06.25; published 25.07.25.

Please cite as:
Anonymous
Peer Review of "Assessment of SARC-F Sensitivity for Probable Sarcopenia Among Community-Dwelling Older Adults: Cross-Sectional Questionnaire Study"
JMIRx Med 2025;6:e78552
URL: https://xmed.jmir.org/2025/1/e78552
doi:10.2196/78552

Peer Review of “Assessment of SARC-F Sensitivity for Probable Sarcopenia Among Community-Dwelling Older Adults: Cross-Sectional Questionnaire Study”

Anonymous

Related Articles:

Companion article: https://www.medrxiv.org/content/10.1101/2023.10.31.23297840v2

Companion article: https://med.jmirx.org/2025/1/e77497

Companion article: https://med.jmirx.org/2025/1/e54475

(JMIRx Med 2025;6:e77582) doi:10.2196/77582

KEYWORDS

sarcopenia; neuromuscular; screening; community; scale; measure; questionnaires; diagnosis; gerontology; older adults; muscular

This is a peer-review report for “Assessment of SARC-F Sensitivity for Probable Sarcopenia Among Community-Dwelling Older Adults: Cross-Sectional Questionnaire Study.”

Round 1 Review

General Comments

This paper [1] conducted a validation to derive a cutoff value that predicted low grip strength from SARC-F (strength, assistance with walking, rising from a chair, climbing stairs, and falls) scores and showed that the cutoff for SARC-F scores is 2 points. Many issues need to be resolved before this study can be published.

Specific Comments

Major Comments

1. The study looked at the association between SARC-F and grip strength, which is not novel. Sarcopenia is poorly defined.
2. The sample size needed to be more adequate, and only 11% of the subjects had lower grip strength.
3. It is acceptable if it is used for estimation or prediction, such as death, but an area under the curve of 0.77 may be too low as an index for diagnosis and discrimination.
4. The Methods describe too few details, and Table 1 provides too little background information.
5. Ultimately, the conclusions that can be drawn from the results should be revised.

Round 2 Review

General Comments

The authors have attempted to revise the manuscript to the best of their ability, but even so, this study seems to lack important points.

Specific Comments

Major Comments

To begin with, SARC-F is a screening indicator for sarcopenia, not for probable sarcopenia (decreased grip strength). If you try to find a cutoff for probable sarcopenia, which is a prestage of sarcopenia, the cutoff value will inevitably be smaller than the cutoff value used to determine sarcopenia. With that in mind, how do you explain the significance of this paper? Please argue the need to screen for decreased grip strength with a cutoff of 2 points rather than screening for sarcopenia with a cutoff of 4 points.

In addition, the cutoff of 2 points on a questionnaire consisting of five items with a range of 0 - 12 points is an extremely low value. The question that arises here is whether there is any point in using this questionnaire in the first place. The authors will first need to show which of the lower-level items contribute strongly to the prediction of grip strength decline as a sensitivity analysis. Then, they should also mention whether the SARC-F should be used as a questionnaire indicator or whether it would be better to use the lower-level items as a new screening indicator.

Minor Comments

Information on ethical matters is lacking.

1. Is there an ethics approval number?
2. It is said that informed consent was not required, but how was information disclosed to the research subjects regarding your research? Was an opt-out notice posted?
3. How was the opportunity for the subjects to decline participation in your research provided?

It says “regularly scheduled physician visits,” but is this study a single or multicenter study?

What is the reason for the subjects’ physician visits? Are the subjects suffering from some disease? If so, the disease

information may be an important confounding factor in this study, so please clearly state the results and show them in Table 1.

Please show the inclusion and exclusion criteria for the subjects.

Who measured grip strength, where, and in what position?

In the Statistical Analysis section, it says “visual histograms,” but they are not shown in the Results. Please show them. In particular, it would be desirable for the histogram of the SARC-F score to be free from extreme bias when conducting the analysis. Please show the histogram for each sex and show that the sampling is appropriate for verifying the value conducted in this study.

Before validating the cutoff value of the SARC-F based on grip strength, it’s crucial to establish a robust relationship between grip strength and the SARC-F. This can be achieved through multiple regression analysis, with grip strength as the dependent variable, the SARC-F as the explanatory variable, and other factors as adjustment factors. This step is essential to ensure the validity of the research.

The factors that may confound the relationship between SARC-F and grip strength have yet to be sufficiently demonstrated. For example, what about cognitive function and physical activity?

The male’s grip strength of 36.3 kg is extremely strong for a subject who should be selected for probable sarcopenia. There is a high possibility of selection bias. Please clearly state in the Discussion how you interpret this point.

As mentioned above, much important information needs to be included, and even though there are limitations from the research planning stage, they should be mentioned in the Discussion.

If you do not present the information mentioned above, please clearly state the limitations of the research in the Discussion section, and also explain why you still think the research results are meaningful and why it is necessary to make the results of this research public.

Conflicts of Interest

None declared.

Reference

1. Propst D, Biscardi L, Dornemann T. Assessment of SARC-F sensitivity for probable sarcopenia among community-dwelling older adults: cross-sectional questionnaire study. JMIRx Med 2025;6:e54475. [doi: 10.2196/54475]

Abbreviations

SARC-F: strength, assistance with walking, rising from a chair, climbing stairs, and falls

Edited by A Schwartz; submitted 15.05.25; this is a non–peer-reviewed article;accepted 15.05.25; published 25.07.25.

Please cite as:
Anonymous
Peer Review of “Assessment of SARC-F Sensitivity for Probable Sarcopenia Among Community-Dwelling Older Adults: Cross-Sectional Questionnaire Study”
JMIRx Med 2025;6:e77582
URL: https://xmed.jmir.org/2025/1/e77582
doi:10.2196/77582

Peer Review of “Cardiotoxicity in Pediatric Cancer Survivorship: Retrospective Cohort Study”

John Lucas Jr

St. Jude Children's Research Hospital, Memphis, TN, United States

Related Articles:

Companion article: https://www.medrxiv.org/content/10.1101/2024.08.10.24311795v1

Companion article: https://med.jmirx.org/2025/1/e79672

Companion article: https://med.jmirx.org/2025/1/e65299

Abstract

(JMIRx Med 2025;6:e79523) doi:10.2196/79523

KEYWORDS

cardiotoxicity; cardiology; cardiovascular; heart; arrhythmias; self-reported questionnaires; oncology; survivors; pediatrics; prevalence; incidence; risk; epidemiology; anthracycline exposure; childhood cancer survivors

This is a peer-review report for “Cardiotoxicity in Pediatric Cancer Survivorship: Retrospective Cohort Study.”

Round 1 Review

Overall Evaluation

This significant and timely manuscript [1], which investigates the long-term cardiovascular complications in pediatric cancer survivors, has notable strengths, including its large cohort size, long-term follow-up, and utilization of a well-established dataset (Childhood Cancer Survivor Study). The methodology is generally sound, and the findings contribute meaningfully to our understanding of cardiotoxicity risks in childhood cancer survivors. However, certain areas necessitate clarification and additional analyses. These are detailed below.

The study relies heavily on self-reported cardiovascular complications, which may introduce reporting bias. While a subset of cases was validated via medical records, the proportion of validated cases is not explicitly stated, and the possibility of underreporting or overreporting remains. The reliance on self-reported cardiovascular complications may have introduced reporting bias into the study. Although some cases were validated through medical records, the proportion of validated cases remains unclear, leaving the potential for underreporting or overreporting. The authors could also consider exploring linkage with external databases (eg, insurance claims, hospital records) for additional validation.

The manuscript presents a risk prediction model (C statistic 0.78), but there is no external validation or discussion of its clinical applicability. Validate the model using an independent dataset (eg, a subset of Childhood Cancer Survivor Study data withheld from model training or another survivor cohort). Report calibration metrics (eg, Hosmer-Lemeshow test, calibration plots) to assess model accuracy. Provide a clinical risk score or decision framework for practical implementation.

The study reports a decreasing risk of cardiotoxicity over time, suggesting improvements in treatment protocols. However, this could be confounded by survivor selection bias (eg, patients with higher early mortality due to severe toxicity were less likely to be included in later eras).

Adjust for potential survivor bias using inverse probability weighting or sensitivity analyses. Consider comparing treatment regimens (eg, changes in anthracycline dosages, cardioprotective measures) across eras to explicitly determine which interventions contributed to reduced risk. The research indicates that the risk of cardiotoxicity diminishes over time, suggesting that treatment protocols have become more effective. However, it is possible that this observation is attributable to survivor selection bias, wherein patients who succumbed to severe toxicity early in the study were not included in subsequent phases. To address potential survivor bias, researchers should employ methodologies such as inverse probability weighting or sensitivity analyses. Additionally, treatment regimens (eg, modifications in anthracycline dosages and cardioprotective measures) should be compared across different time periods to ascertain which interventions are responsible for the diminished risk.

The study focuses on clinically evident cardiovascular complications but does not assess subclinical cardiotoxicity, which could be detected via biomarkers or imaging.

Incorporate cardiac biomarkers (eg, troponins, N-terminal pro–brain natriuretic peptide) in a subset of survivors to identify early signs of myocardial damage. Perform echocardiographic or cardiac magnetic resonance imaging evaluations in a subgroup to detect preclinical cardiac dysfunction. This could strengthen the study's ability to recommend early intervention strategies.

The authors appropriately point out the opportunity to improve early intervention by identifying a subset of survivors for early myocardial damage using cardiac biomarkers and imaging. While this is not possible in the present study, future studies incorporating this approach would allow for detection of subclinical cardiotoxicity.

The manuscript discusses risk factors but does not evaluate protective factors (eg, exercise, angiotensin-converting enzyme inhibitors, β-blockers). Analyze whether lifestyle modifications (eg, regular exercise) or cardioprotective medications influence the incidence of cardiotoxicity. Conduct a subgroup analysis on survivors who received cardioprotective interventions versus those who did not.

Please indicate whether the proportional hazards assumptions were tested and consider reporting Schoenfeld residuals or time-dependent covariate analyses.

Please include more details on how missing data were handled.

Were there particular domains of quality of life that were lower among those with cardiovascular complications?

Consider adding detailed figure legends to improve readability and refining axis labels in existing figures.

A table summarizing key risk factors with adjusted hazard ratios and *P* values would be beneficial.

Round 2 Review

Please state the proportion of cases with cardiovascular events confirmed by medical record review.

Please discuss the increased cardiotoxicity observed in male survivors. Was this due to treatment or other comorbidities that exacerbated previously subclinical cardiac exposures?

Please provide a thoughtful description of how the risk model could be integrated into previously described models and recommendations for cardiac risk groups like the International Late Effects of Childhood Cancer Guideline Harmonization Group.

Please standardize the reporting/formatting for data into a table format more typical for manuscript reporting for complication rates, multivariate cox regression, and temporal trends.

Please provide a table or figure for the treatment era analysis.

Please provide a table or figure for the sibling controls comparison. Is this after adjustment for age, gender, etc?

The CI of cardiovascular complications in childhood cancer survivors data is shown in a nonstandard stacked bar plot format. Please show as CI curves.

Conflicts of Interest

None declared.

Reference

1. Mansoor M, Ibrahim A. Cardiotoxicity in Pediatric Cancer Survivorship: Retrospective Cohort Study. JMIRx Med 2025;6:e65299. [doi: 10.2196/65299]

Edited by F Wu; submitted 23.06.25; this is a non–peer-reviewed article;accepted 23.06.25; published 31.07.25.

Please cite as:
Lucas Jr J
Peer Review of "Cardiotoxicity in Pediatric Cancer Survivorship: Retrospective Cohort Study"
JMIRx Med 2025;6:e79523
URL: https://xmed.jmir.org/2025/1/e79523
doi:10.2196/79523

Peer Review of “Cardiotoxicity in Pediatric Cancer Survivorship: Retrospective Cohort Study”

Archana Adhikari

Sikkim Manipal University, Gangtok, India

Related Articles:

Companion article: https://www.medrxiv.org/content/10.1101/2024.08.10.24311795v1

Companion article: https://med.jmirx.org/2025/1/e79672

Companion article: https://med.jmirx.org/2025/1/e65299

Abstract

(JMIRx Med 2025;6:e79521) doi:10.2196/79521

KEYWORDS

cardiotoxicity; cardiology; cardiovascular; heart; arrhythmias; self-reported questionnaires; oncology; survivors; pediatrics; prevalence; incidence; risk; epidemiology; anthracycline exposure; childhood cancer survivors

This is a peer-review report for “Cardiotoxicity in Pediatric Cancer Survivorship: Retrospective Cohort Study.”

Round 1 Review

General Comments

This paper [1] gives valuable insights into cardiotoxicity in pediatric cancer survivorship: patterns, predictors, and implications for long-term care. The results and methodology are sound. However, some minor revisions would improve clarity and strengthen the overall impact of this paper. Below are my suggestions.

Major Comments

1. Method section (study population and data source): In the Method section, specifically the fourth line, the description “of 21 at one of 31 participating institutes” is unclear. The sentence should be revised for better clarity.
2. Missing answer for seventh objective: The answer to the seventh objective is unclear.

Minor Comments

1. Result presentation: It would be better if the results were presented in tabular format for easier comprehension. A table would help summarize the key findings and increase readability.
2. Clarity in results numbering: To improve clarity, it would be beneficial to present all the results with corresponding numbers, matching each result with the respective objective number for easier reference and alignment.

Conflicts of Interest

None declared.

Reference

1. Mansoor M, Ibrahim A. Cardiotoxicity in Pediatric Cancer Survivorship: Retrospective Cohort Study. JMIRx Med 2025;6:e65299. [doi: 10.2196/65299]

Edited by F Wu; submitted 23.06.25; this is a non–peer-reviewed article;accepted 23.06.25; published 31.07.25.

Please cite as:
Adhikari A
Peer Review of "Cardiotoxicity in Pediatric Cancer Survivorship: Retrospective Cohort Study"
JMIRx Med 2025;6:e79521
URL: https://xmed.jmir.org/2025/1/e79521
doi:10.2196/79521

Peer Review of “Willingness to Pay for the COVID-19 Vaccine and Its Correlates in Bangladesh: Cross-Sectional Study”

Enamul Hoque

The University of Western Australia, 35 Stirling Highway, Perth, Australia

Related Articles:

Companion article: https://www.medrxiv.org/content/10.1101/2024.12.03.24318442v1

Companion article: https://med.jmirx.org/2025/1/e79352

Companion article: https://med.jmirx.org/2025/1/e69827

(JMIRx Med 2025;6:e79354) doi:10.2196/79354

KEYWORDS

Bangladesh; willingness to pay; vaccines; COVID-19; infectious diseases; infection control; public health; public safety; cross-sectional study; financial

This is a peer-review report for “Willingness to Pay for the COVID-19 Vaccine and Its Correlates in Bangladesh: Cross-Sectional Study.”

Round 1 Review

Minor Comments

1. In lines 79 and 80 of the manuscript [1], it is confusing why this wouldn’t be considered nationally representative if the data collection was conducted online.
2. As around 50% of the people are not interested in paying for the vaccine, this result should be considered with caution.

Conflicts of Interest

None declared.

Reference

1. Hossain MB, Alam MZ, Islam MS, et al. Willingness to pay for the COVID-19 vaccine and its correlates in Bangladesh: cross-sectional study. JMIRx Med 2025;6:e69827. [doi: 10.2196/69827]

Edited by F Wu; submitted 19.06.25; this is a non–peer-reviewed article;accepted 19.06.25; published 15.08.25.

Please cite as:
Hoque E
Peer Review of “Willingness to Pay for the COVID-19 Vaccine and Its Correlates in Bangladesh: Cross-Sectional Study”
JMIRx Med 2025;6:e79354
URL: https://xmed.jmir.org/2025/1/e79354
doi:10.2196/79354

Peer Review of "Willingness to Pay for the COVID-19 Vaccine and Its Correlates in Bangladesh: Cross-Sectional Study"

Enamul Kabir

University of Southern Queensland, UniSQ Toowoomba, 487-535 West St, Queensland, Australia

Related Articles:

Companion article: https://www.medrxiv.org/content/10.1101/2024.12.03.24318442v1

Companion article: https://med.jmirx.org/2025/1/e79352

Companion article: https://med.jmirx.org/2025/1/e69827

(JMIRx Med 2025;6:e79355) doi:10.2196/79355

KEYWORDS

Bangladesh; willingness to pay; vaccines; COVID-19; infectious diseases; infection control; public health; public safety; cross-sectional study; financial

This is a peer-review report for "Willingness to Pay for the COVID-19 Vaccine and Its Correlates in Bangladesh: Cross-Sectional Study."

Round 1 Review

This paper [1] addresses an important and timely topic—willingness to pay for COVID-19 vaccines in a developing country context. Understanding willingness to pay is essential not only for informing current vaccine financing strategies but also for shaping policies related to equitable vaccine access in response to future public health challenges. The study is well-conceived and provides valuable insights into vaccine affordability and public perception in Bangladesh. With some refinements in presentation, statistical interpretation, and policy framing, the paper will be well-positioned for publication.

The abstract would benefit from being more concise and should more clearly highlight the key policy implications of the findings. Additionally, the statistical interpretation of the adjusted odds ratios (aORs) requires careful attention. Several aORs are reported with values close to 1 (eg, family income aOR 1.0, *P*=.039; vaccine knowledge aOR 1.1, *P*=.003; behavioral practices aOR 1.1, *P*<.001), suggesting minimal effect sizes, yet they are statistically significant. While such significance may be driven by the large sample size, reporting CIs would allow for a more meaningful interpretation of the strength and direction of these associations.

The paper would also benefit from greater clarity around the construction of variables and the underlying measurement models. It is unclear how multiple survey items were combined to form factors such as knowledge, attitudes, and behavioral constructs. Using exploratory factor analysis could be beneficial to validate the grouping of items into coherent factors and strengthen construct validity. Providing factor loadings or at least a brief description of the item-grouping process would enhance the methodological transparency of the study.

Another area for improvement involves the reporting of the income variable. In both Table 1 and Table 2, income appears to be modeled as a continuous variable, but the unit of measurement is not specified. Without this information, it is difficult to interpret an odds ratio of 1.0 meaningfully. If income is measured in small units (eg, Bangladeshi taka), the impact of each unit increase would be negligible. Categorizing income into meaningful brackets (eg, low, middle, high) and using those categories in logistic regression would make the results more interpretable and policy relevant. Additionally, the CIs for some variables in Table 2—such as income and COVID-19 vaccine conspiracy beliefs—appear to suggest nonsignificance, yet they are reported as significant. This inconsistency should be carefully reviewed and clarified.

Some of the measured constructs, such as knowledge and perceived susceptibility, show relatively low internal consistency (eg, Cronbach α of 0.612 and 0.657, respectively). It would be helpful for the authors to explain why these values are considered acceptable in this context or to discuss efforts made to improve reliability through item refinement or scale revision. Furthermore, the combination of nonprobability online sampling and quota sampling should be more clearly justified. While practical during a pandemic, it raises concerns about representativeness and potential sampling bias, which should be acknowledged more explicitly in the Discussion.

The manuscript would also benefit from a thorough review for minor language and formatting issues. For instance, the phrase "explains explains" on page 13 should be corrected. Variable labels and descriptions in tables should be presented clearly and consistently.

Overall, this is a valuable study that contributes to the growing body of literature on COVID-19 vaccine access and health

economics. With revisions to enhance clarity, statistical reporting, and methodological transparency, the paper has strong potential for publication and meaningful policy impact.

Conflicts of Interest

None declared.

Reference

1. Hossain MB, Alam MZ, Islam MS, et al. Willingness to pay for the COVID-19 vaccine and its correlates in Bangladesh: cross-sectional study. JMIRx Med 2025;6:e69827. [doi: 10.2196/69827]

Abbreviations

aOR: adjusted odds ratio

Edited by F Wu; submitted 19.06.25; this is a non–peer-reviewed article;accepted 19.06.25; published 15.08.25.

Please cite as:
Kabir E
Peer Review of "Willingness to Pay for the COVID-19 Vaccine and Its Correlates in Bangladesh: Cross-Sectional Study"
JMIRx Med 2025;6:e79355
URL: https://xmed.jmir.org/2025/1/e79355
doi:10.2196/79355

Peer Review of “Willingness to Pay for the COVID-19 Vaccine and Its Correlates in Bangladesh: Cross-Sectional Study”

Jatina Vij

Post Graduate Institute of Medical Education & Research, Chandigarh, Madhya Marg, Sector 12, Chandigarh, India

Related Articles:

Companion article: https://www.medrxiv.org/content/10.1101/2024.12.03.24318442v1

Companion article: https://med.jmirx.org/2025/1/e79352

Companion article: https://med.jmirx.org/2025/1/e69827

(JMIRx Med 2025;6:e79353) doi:10.2196/79353

KEYWORDS

Bangladesh; willingness to pay; vaccines; COVID-19; infectious diseases; infection control; public health; public safety; cross-sectional study; financial

This is a peer-review report for “Willingness to Pay for the COVID-19 Vaccine and Its Correlates in Bangladesh: Cross-Sectional Study.”

Round 1 Review

General Comments

This paper [1] examines willingness to pay (WTP) for COVID-19 vaccines in Bangladesh using a cross-sectional survey. The integration of the health belief model and theory of planned behavior adds a theoretical foundation to the analysis. The study is well-structured, and the use of hierarchical logistic regression strengthens its analytical rigor.

However, several issues need to be addressed before acceptance. The sampling methodology raises concerns about representativeness, particularly due to the mix of online and face-to-face data collection. Additionally, some statistical interpretations require further clarification, and the discussion on policy implications could be expanded to provide actionable recommendations. Addressing these concerns will enhance the overall impact and credibility of the study.

Specific Comments

Major Comments

1. The study employs both online and face-to-face data collection. However, the online survey may have overrepresented educated and tech-savvy individuals, while the face-to-face survey followed quota sampling.

2. Clarify the adjusted odds ratio interpretation. Some adjusted odds ratio values are close to 1, making practical significance questionable.

3. The impact of administrative divisions (eg, Sylhet having 4× higher WTP) should be further discussed. Are these differences due to economic, cultural, or policy variations?

4. While the study suggests subsidized vaccination programs, it would be helpful to compare findings with other low- and middle-income countries’ WTP trends.

Minor Comments

5. Ensure table captions clearly describe what is presented (eg, Table 2 should explicitly state that it presents logistic regression results).

6. Some sections contain grammatical errors and awkward phrasing (eg, “knowledge about the vaccine, vaccine process, conspiracy beliefs, behavioral practice, attitude toward a vaccine”; this list is repetitive and unclear).

Conflicts of Interest

None declared.

Reference

1. Hossain MB, Alam MZ, Islam MS, et al. Willingness to pay for the COVID-19 vaccine and its correlates in Bangladesh: cross-sectional study. JMIRx Med 2025;6:e69827. [doi: 10.2196/69827]

Abbreviations

WTP: willingness to pay

Edited by F Wu; submitted 19.06.25; this is a non–peer-reviewed article;accepted 19.06.25; published 15.08.25.

Please cite as:
Vij J
Peer Review of "Willingness to Pay for the COVID-19 Vaccine and Its Correlates in Bangladesh: Cross-Sectional Study"
JMIRx Med 2025;6:e79353
URL: https://xmed.jmir.org/2025/1/e79353
doi:10.2196/79353

Peer Review of “Effect of Thermal and Vibration Changes on Automated External Defibrillator Circuit Boards: Finite Element Analysis Study”

Anonymous

Related Articles:

Companion article: https://arxiv.org/abs/2309.14747v1

Companion article: https://med.jmirx.org/2025/1/e80135

Companion article: https://med.jmirx.org/2025/1/e53208

Abstract

(JMIRx Med 2025;6:e80137) doi:10.2196/80137

KEYWORDS

circuit board; automated external defibrillator; heart; cardiology; vibration; thermal changes; medical devices

This is a peer-review report for “Effect of Thermal and Vibration Changes on Automated External Defibrillator Circuit Boards: Finite Element Analysis Study.”

Round 1 Review

General Comments

I have read the submitted manuscript [1] to your journal entitled “Analysis of Vibration and Thermal of a Modeled Circuit Board of Automated External Defibrillator (AED) Medical Device.” The author utilized finite element analysis to simulate static and dynamic testings in determining the vibration and thermal effects on the operations of the automated external defibrillator medical device.

Based on the outcomes of this conducted study and the potential benefits from this study, I would recommend this manuscript to be published in your reputable journal after the minor comments are properly addressed.

Specific Comments

Minor Comments

There are 4-member and 8-member supports. The author should provide more evidence on how these affect the analysis results.

This manuscript will benefit from other works on the failure modes of materials. The author is encouraged to expand the Literature Review section and also provide more references.

The grammar should be refined. Some of the grammar in this manuscript should be corrected. For example, “This study was performed to analysis the vibration...” “Analyse.” Fig 14 = Fig. 14 or Figure 14. The unit of temperature is degree c; the *c* is capitalized.

When citing a research paper within the manuscript, it is the first/lead author who should be named in the “Name et al” format; the author should use a consistent citation format.

The author should remove parentheses from the manuscript title: “Analysis of Vibration and Thermal of a Modelled Circuit Board of Automated External Defibrillator Medical Device”

For reproducibility by other researchers, the author should consider providing simulation data as supplementary information.

The author should properly cite other works where applicable. For example, “From other research results, it can be verified that the natural frequencies...” The author needs to insert appropriate citations for comments like these.

Conflicts of Interest

None declared.

Reference

1. Olalere SO. Effect of Thermal and Vibration Changes on Automated External Defibrillator Circuit Boards: Finite Element Analysis Study. JMIRx Med 2025;6:e53208. [doi: 10.2196/53208]

Edited by T Leung; submitted 04.07.25; this is a non–peer-reviewed article;accepted 04.07.25; published 19.08.25.

Please cite as:
Anonymous
Peer Review of "Effect of Thermal and Vibration Changes on Automated External Defibrillator Circuit Boards: Finite Element Analysis Study"
JMIRx Med 2025;6:e80137
URL: https://xmed.jmir.org/2025/1/e80137
doi:10.2196/80137

Peer Review of "Effect of Thermal and Vibration Changes on Automated External Defibrillator Circuit Boards: Finite Element Analysis Study"

Ayobami Akinfenwa, MSc

Related Articles:

Companion article: https://arxiv.org/abs/2309.14747v1

Companion article: https://med.jmirx.org/2025/1/e80135

Companion article: https://med.jmirx.org/2025/1/e53208

Abstract

(JMIRx Med 2025;6:e80142) doi:10.2196/80142

KEYWORDS

circuit board; automated external defibrillator; heart; cardiology; vibration; thermal changes; medical devices

This is a peer-review report for "Effect of Thermal and Vibration Changes on Automated External Defibrillator Circuit Boards: Finite Element Analysis Study."

Round 1 Review

General Comments

This paper [1] considered the vibration and thermal analysis of a modeled circuit board of an automated external defibrillator (AED) using Ansys. The vibration failure in the modeled circuit board with four rigid supports was pronounced starting at the taller components, including the capacitor. The board was reinforced to an 8–rigid support system, reducing the failure around the rigid supports. The thermal failure started from the battery position, causing thermal dissipation to other parts of the board, ultimately leading to the failure of the circuit board and the AED.

Specific Comments

Major Comments

1. "The goal is to analyse the effect of vibration and thermal experience on the AED based on its operation." Is this your research statement? I suppose the topic suggests you analyzed the modeled circuit board, not the overall AED system. If not, how did you measure the overall effect on the AED?

2. The author may also need to discuss the importance of the circuit boards in an AED in the Introduction.

3. The figures will need a little more discussion.

Minor Comments

4. "Vibration and Thermal Analysis on Modeled Circuit Board of Automated External Defibrillator (AED) Medical Device" will likely communicate the title better.

5. I find it a bit difficult to understand this line: "Fatigue failure under sinusoidal vibration loading for component by comparing the vibration failure test, FEA, and theoretical test (Y.S.Chen, 2008)." What did you want to say?

6. Figure 1 will need relabeling. The labeling seems to cover some parts of the board. A transparent background could help.

Round 2 Review

1. The recommendation about explaining how the author measured the overall effect of the analysis on the AED or the board and the recommendation that the author should provide more evidence on how the 4-member and 8-member supports affect the analysis result have not been answered or addressed in the manuscript. The author should consider these.

2. The author will also need to be consistent. Is Figure 5 the same as Figure 5 or Fig. 5? It should be corrected for all other instances.

3. Additional citations might be needed in the work; it still looks like over 40% of the citations are 15 years or older. Also, "et al.," should be in italics with a period after the "al." The Discussion also seems not well discussed in relations to previous works.

4. The template format should also be considered carefully.

Conflicts of Interest

None declared.

Reference

1. Olalere SO. Effect of Thermal and Vibration Changes on Automated External Defibrillator Circuit Boards: Finite Element Analysis Study. JMIRx Med 2025;6:e53208. [doi: 10.2196/53208]

Abbreviations

AED: automated external defibrillator

Edited by T Leung; submitted 04.07.25; this is a non–peer-reviewed article;accepted 04.07.25; published 19.08.25.

Please cite as:
Akinfenwa A
Peer Review of "Effect of Thermal and Vibration Changes on Automated External Defibrillator Circuit Boards: Finite Element Analysis Study"
JMIRx Med 2025;6:e80142
URL: https://xmed.jmir.org/2025/1/e80142
doi:10.2196/80142

Peer Review of "Challenges in Implementing a Mobile AI Chatbot Intervention for Depression Among Youth on Psychiatric Waiting Lists: Randomized Controlled Study Termination Report"

Maria da Graca Ambrosio

University of Oxford, Oxford, United Kingdom

Related Articles:

Companion article: https://www.medrxiv.org/content/10.1101/2024.11.25.24317880v1

Companion article: https://med.jmirx.org/2025/1/e82083

Companion article: https://med.jmirx.org/2025/1/e70960

(JMIRx Med 2025;6:e82073) doi:10.2196/82073

KEYWORDS

randomized controlled trial; AI chatbot; acceptance and commitment therapy; mental health; psychiatry; children; adolescents; Japan

This is a peer-review report for "Challenges in Implementing a Mobile AI Chatbot Intervention for Depression Among Youth on Psychiatric Waiting Lists: Randomized Controlled Study Termination Report."

Round 1 Review

General Comments

This paper [1] describes the results of a parallel group randomized controlled trial that examined the feasibility of an artificial intelligence (AI) chatbot-led mental health intervention to support pediatric patients on the psychiatry waitlists in Japan. The article is well-written and organized, and the objectives of the study are clearly stated. Methodology elements such as eligibility criteria, information sources, and data collection process are clear. A clear list of outcomes and variables for which data were researched is presented. The authors provide an important contribution to the field by reporting on factors that challenge adolescents' engagement in digital mental health interventions and providing meaningful recommendations for future research.

Specific Comments

Major Comments

1. How many chatbots were shortlisted, and why was emol favored over the others, given the selection criteria? (Under AI Chatbot Selection Process.)

2. How are the six core processes of acceptance and commitment therapy delivered in the AI chatbot (under Intervention Group)? Expand more on each section. How does the session meet the core processes of acceptance and commitment therapy—acceptance, cognitive defusion, being present, self as context, values, and committed action?

3. How was the section structured? Did adolescents go through modules? Could they write anything to the chatbot, or was the content predefined? Were the sessions sequentially delivered or not? Could they access previously completed modules or track their progress?

4. Were there any safeguarding links and referral contacts built into the chatbot in case participants needed additional support beyond those offered by the chatbot? If yes, I recommend including it under the ethics paragraph.

5. How were you planning to investigate engagement? Would you report on the frequency of use, number of interactions with the chatbot, or amount of content visualized by participants? Even though the study's main questions are not focused on engagement, I suggest that the authors consider including an engagement outcome paragraph right after the secondary outcomes.

Minor Comments

6. I recommend moving all hyperlinks to the appendix and including an image of the chatbot. I also recommend that authors include an image of the intervention delivered through the hospital website.

7. Please state the statistical methods used to deal with missing data.

8. In the Discussion, you argue that young people prefer online mental health support over in-person support [2]. I believe you could discuss this a bit more in your Introduction paragraph to strengthen your discussion regarding the potential gap online services could fill.

9. I recommend including a paragraph under the Introduction on previous Japanese studies focusing on chatbot-led or digital

mental/public health interventions to provide an overview of the current population uptake of digital health interventions.

Conflicts of Interest

None declared.

References

1. Fujita J, Yano Y, Shinoda S, et al. Challenges in implementing a mobile AI chatbot intervention for depression among youth on psychiatric waiting lists: randomized controlled study termination report. JMIRx Med 2025;6:e70960. [doi: 10.2196/70960]
2. Rickwood DJ, Mazzer KR, Telford NR. Social influences on seeking help from mental health services, in-person and online, during adolescence and young adulthood. BMC Psychiatry 2015 Mar 7;15:40. [doi: 10.1186/s12888-015-0429-6] [Medline: 25886609]

Abbreviations

AI: artificial intelligence

Edited by S Amal; submitted 08.08.25; this is a non–peer-reviewed article;accepted 08.08.25; published 05.09.25.

Please cite as:
Ambrosio MDG
Peer Review of "Challenges in Implementing a Mobile AI Chatbot Intervention for Depression Among Youth on Psychiatric Waiting Lists: Randomized Controlled Study Termination Report"
JMIRx Med 2025;6:e82073
URL: https://xmed.jmir.org/2025/1/e82073
doi:10.2196/82073

Peer Review of "Challenges in Implementing a Mobile AI Chatbot Intervention for Depression Among Youth on Psychiatric Waiting Lists: Randomized Controlled Study Termination Report"

Edel Ennis

University of Ulster, School of Psychology, Coleraine, United Kingdom

Related Articles:

Companion article: https://www.medrxiv.org/content/10.1101/2024.11.25.24317880v1

Companion article: https://med.jmirx.org/2025/1/e82083

Companion article: https://med.jmirx.org/2025/1/e70960

(JMIRx Med 2025;6:e82071) doi:10.2196/82071

KEYWORDS

randomized controlled trial; AI chatbot; acceptance and commitment therapy; mental health; psychiatry; children; adolescents; Japan

This is a peer-review report for "Challenges in Implementing a Mobile AI Chatbot Intervention for Depression Among Youth on Psychiatric Waiting Lists: Randomized Controlled Study Termination Report."

Round 1 Review

Thank you for inviting me to review this paper [1]. However, my suggestion would be that this paper should be rejected. I am very cognizant of publication biases, and I am a firm believer that the publication of negative results is very important. I therefore have no problem with the fact that the sample decreased to zero. However, I do believe that more detail is needed in terms of *why* people disengaged. The rationale for the paper is set up as efficacy of the intervention, but the main message of the paper is that the sample declined. I would therefore like more emphasis on qualitative interviews that examined why people disengaged. Follow-up work such as this would make a very interesting paper.

Conflicts of Interest

None declared.

Reference

1. Fujita J, Yano Y, Shinoda S, et al. Challenges in implementing a mobile AI chatbot intervention for depression among youth on psychiatric waiting lists: randomized controlled study termination report. JMIRx Med 2025;6:e70960. [doi: 10.2196/70960]

Edited by S Amal; submitted 08.08.25; this is a non–peer-reviewed article;accepted 08.08.25; published 05.09.25.

Please cite as:
Ennis E
Peer Review of "Challenges in Implementing a Mobile AI Chatbot Intervention for Depression Among Youth on Psychiatric Waiting Lists: Randomized Controlled Study Termination Report"
JMIRx Med 2025;6:e82071
URL: https://xmed.jmir.org/2025/1/e82071
doi:10.2196/82071

Peer Review of "Challenges in Implementing a Mobile AI Chatbot Intervention for Depression Among Youth on Psychiatric Waiting Lists: Randomized Controlled Study Termination Report"

Beatrice Tosti

University of Cassino and Southern Lazio, Cassino, Italy

Related Articles:

Companion article: https://www.medrxiv.org/content/10.1101/2024.11.25.24317880v1

Companion article: https://med.jmirx.org/2025/1/e82083

Companion article: https://med.jmirx.org/2025/1/e70960

(JMIRx Med 2025;6:e82074) doi:10.2196/82074

KEYWORDS

randomized controlled trial; AI chatbot; acceptance and commitment therapy; mental health; psychiatry; children; adolescents; Japan

This is a peer-review report for "Challenges in Implementing a Mobile AI Chatbot Intervention for Depression Among Youth on Psychiatric Waiting Lists: Randomized Controlled Study Termination Report."

Round 1 Review

General Comments

The topic and objectives of the study [1] are certainly interesting, as depression among young individuals is an increasingly pervasive and growing problem globally, exacerbated by the COVID-19 pandemic, as the authors themselves point out. Furthermore, the use of artificial intelligence to support traditional methods of treating this condition makes the study topical. The paper is well-written and comprehensible throughout; the supporting bibliography is adequate; it has a good methodological approach, with clear and well-defined objectives, and an accurate description of the inclusion and exclusion criteria for participants. Although the statistical analyses planned by the authors are consistent with the objectives they have defined, the lack of availability of data on which to carry out these analyses and, therefore, the absence of results does not allow an evaluation of this specific aspect. However, the authors have posited potential explanations for instances of nonadherence to the intervention protocol, which are substantiated by extant literature on the subject, therefore apprising the reader of the possible limitations of this type of intervention in this specific population that fulfills certain inclusion criteria. The paper thus provides a cue and guidance for future studies in this field. Lastly, as stated in the major comments below, the major shortcoming of this study is the lack of clarity as to whether the authors used an active or nonactive control group.

Specific Comments

Major Comments

1. In the Study Design paragraph, the authors stated that the control group would receive standard care (making it an active control group), while in the Control Group paragraph, they stated that they would receive general mental health information and would undergo online evaluations and diary recordings (making it a nonactive control group). It is not clear if the authors deem these two procedures similar. In the event that they do not regard them as analogous, it would be beneficial to ascertain which of the two would have been delivered to the control group. Furthermore, it would be appreciated if the authors could provide an explanation and make the appropriate adjustments in the manuscript about (1) what standard care would have comprised and (2) what is the nature of the short video programs that participants received as general mental health information, in order to enable the reader to ascertain whether they are informational videos, mental health support videos, etc.

Round 2 Review

General Comments

I would like to express my gratitude to the authors for implementing the requested revisions, which have served to enhance the clarity and thoroughness of the manuscript. Still, there are some elements that, in my view, would benefit from modification.

Specific Comments

Major Comments

1. Supplementary Table 1 and the supplementary figure are missing.
2. The sentence “AI chatbot emol features a friendly character name ‘Roku’” is redundant, as the same concept is repeated in the preceding sentence (in the AI Chatbot Selection Process paragraph).
3. The following sentence is repeated twice: “Weekly online assessments were conducted at Week 0, during the intervention period, and at Week 9” (in the Intervention Group paragraph).
4. The sentence “Non physician research assistants encouraged participants to use the pen consistently for their diary entries and performed minimal mental status checks during these assessments” is redundant, as the same concept is repeated afterward in the same paragraph (Intervention Group section). Therefore, it should be deleted to streamline the text.
5. In what manner was the viewing of the videos organized for the control group? Was a schedule in place, or were the participants free to watch the videos at their own discretion? Furthermore, how was the actual viewing of the videos ascertained?
6. In my personal view, the use of an active control group would have been a valuable approach, for instance, by comparing two distinct chatbots providing different types of therapy, the evaluation of which would have determined which one would prove to be more efficacious in terms of symptoms improvement. This approach would have ensured that both groups received a therapeutic intervention and could have provided additional information in terms of engagement and usability. The authors stated that the design they chose “reflects the real-world experience of many psychiatric waiting list patients in Japan,” but as they also declared, “the lack of timely intervention can exacerbate symptoms and increase the risk of severe outcomes.” Therefore, given such a risk, my question is: what is the rationale behind the authors’ decision to employ a passive control group?
7. The concept expressed in the sentence “Another patient refused participation due to concerns about the diary entry, and the third patient was excluded after starting therapy at another facility” is also conveyed in the preceding sentence (in the Results paragraph). It is recommended that one of the two sentences be deleted.

Conflicts of Interest

None declared.

Reference

1. Fujita J, Yano Y, Shinoda S, et al. Challenges in implementing a mobile AI chatbot intervention for depression among youth on psychiatric waiting lists: randomized controlled study termination report. JMIRx Med 2025;6:e70960. [doi: 10.2196/70960]

Edited by S Amal; submitted 08.08.25; this is a non–peer-reviewed article;accepted 08.08.25; published 05.09.25.

Please cite as:
Tosti B
Peer Review of “Challenges in Implementing a Mobile AI Chatbot Intervention for Depression Among Youth on Psychiatric Waiting Lists: Randomized Controlled Study Termination Report”
JMIRx Med 2025;6:e82074
URL: https://xmed.jmir.org/2025/1/e82074
doi:10.2196/82074

Peer Review of “Perception and Impact of White Spot Lesions in Young People Undergoing Orthodontic Treatment and Their Guardians: Protocol for a Mixed Methods Study”

Jonathan Shaw, BS

School of Medicine, California University of Science and Medicine, Colton, CA, United States

Related Articles:

Companion article: http://preprints.jmir.org/preprint/60213

Companion article: https://med.jmirx.org/2025/1/e80139

Companion article: https://med.jmirx.org/2025/1/e60213

(JMIRx Med 2025;6:e80143) doi:10.2196/80143

KEYWORDS

orthodontics; white spot lesions; fixed appliances; dentistry

This is a peer-review report for “Perception and Impact of White Spot Lesions in Young People Undergoing Orthodontic Treatment and Their Guardians: Protocol for a Mixed Methods Study.”

Round 1 Review

General Comments

This paper [1] appears to me to be well-written and adequately cited. I believe that this paper will contribute to the literature once the study commences and the data are collected and analyzed. However, I do have some questions/concerns regarding the study design and potential data analysis that I have included in my comments below. I would like the authors of this paper to review these comments/recommendations and to either implement them as they see fit or justify why they believe they do not need to.

Specific Comments

Major Comments

1. Line 170, you state that this study is purely descriptive, so a power analysis is not required. How will you control for confounding variables such as cultural beliefs which may be over- or underrepresented in your participant pool? Additionally, how will you ensure that your participant demographics allow for the generalization of this paper’s findings to patient populations outside of Liverpool?

2. Line 203, you mention that sampling will be based on age, gender, ethnicity, etc. However, Table 2 does not mention ethnicity. Could you edit Table 2 to mention ethnicity or edit Line 203 to remove ethnicity. I would recommend editing the table because I believe the participant demographics to be important, especially since different cultures may approach esthetics and health beliefs differently. This concern regarding culture connects with major comment 1.

Minor Comments

3. Line 129, “Sponsorship will be sort from...,” please change to “Sponsorship will be sought from…”

4. Line 167, you state that a sample size of 200 respondents is sufficient for Part 1 of the study. Could you justify this estimate in a more thorough way other than stating that it is a “pragmatic estimation?”

5. Line 173, you state that participants will be contacted on the same day as their orthodontic appointment. Will this be before or after the appointment? Will participants be compensated for their time? How will you ensure that participants’ rights are respected and that they do not feel pressured into participating?

6. Line 427, “or childs name?” please change to “or child’s name?”

Conflicts of Interest

None declared.

Reference

1. Hassan AO, Doughty J, Harrison J. Perception and impact of white spot lesions in young people undergoing orthodontic treatment and their guardians: protocol for a mixed methods study. JMIRx Med 2025;6:e60213. [doi: 10.2196/60213]

Edited by E Meinert, T Leung; submitted 04.07.25; this is a non–peer-reviewed article;accepted 04.07.25; published 12.09.25.

Please cite as:
Shaw J
Peer Review of "Perception and Impact of White Spot Lesions in Young People Undergoing Orthodontic Treatment and Their Guardians: Protocol for a Mixed Methods Study"
JMIRx Med 2025;6:e80143
URL: https://xmed.jmir.org/2025/1/e80143
doi:10.2196/80143

Peer Review of “Perception and Impact of White Spot Lesions in Young People Undergoing Orthodontic Treatment and Their Guardians: Protocol for a Mixed Methods Study”

Abdolreza Jamilian

The City of London Dental School, University of Greater Manchester, Bolton, United Kingdom

Related Articles:

Companion article: http://preprints.jmir.org/preprint/60213

Companion article: https://med.jmirx.org/2025/1/e80139

Companion article: https://med.jmirx.org/2025/1/e60213

Abstract

(JMIRx Med 2025;6:e80140) doi:10.2196/80140

KEYWORDS

orthodontics; white spot lesions; fixed appliances; dentistry

This is a peer-review report for “Perception and Impact of White Spot Lesions in Young People Undergoing Orthodontic Treatment and Their Guardians: Protocol for a Mixed Methods Study.”

Round 1 Review

1. The abstract of this paper [1] must be revised. “Several studies explore the prevention and/or treatment of WSL” is inappropriate for the abstract.
2. The Methods section describes the mixed methods approach well, but the recruitment process could be elaborated. For example, how will convenience sampling be conducted to avoid bias? Including qualitative and quantitative data is well-justified, but there is no mention of how the two datasets will be integrated into the analysis. More details on how the qualitative data will expand upon the quantitative findings would strengthen the methodology.
3. The sample size rationale is explained well, though stating why a power analysis is unnecessary for a descriptive study could prevent confusion.
4. In the Results section, it would be useful to clarify how data from questionnaires and interviews will be compared and whether there is an expectation of divergence between parent and child responses.
5. The Limitations section acknowledges some important aspects, such as recruiting from only one hospital, but it does not address potential biases in self-reported data. There is also no mention of how the study will address participants’ potential reluctance to report negative experiences due to social desirability bias. Expanding on these limitations and how the study will mitigate them would improve transparency.
6. The potential psychological impact of white spot lesions could be expanded upon in the Discussion, especially regarding how white spot lesions may affect patient compliance and satisfaction posttreatment.
7. To expand the Discussion, the following article must be cited: Jamloo H, Majidi K, Noroozian N, et al. Effect of fluoride on preventing orthodontics treatments-induced white spot lesions: an umbrella meta-analysis. *Clin Investig Orthod*. April 19, 2024;83(2):53 - 60. [doi: 10.1080/27705781.2024.2342732]

Conflicts of Interest

None declared.

Reference

1. Hassan AO, Doughty J, Harrison J. Perception and impact of white spot lesions in young people undergoing orthodontic treatment and their guardians: protocol for a mixed methods study. JMIRx Med 2025;6:e60213. [doi: 10.2196/60213]

Edited by E Meinert, T Leung; submitted 04.07.25; this is a non–peer-reviewed article;accepted 04.07.25; published 12.09.25.

Please cite as:
Jamilian A
Peer Review of "Perception and Impact of White Spot Lesions in Young People Undergoing Orthodontic Treatment and Their Guardians: Protocol for a Mixed Methods Study"
JMIRx Med 2025;6:e80140
URL: https://xmed.jmir.org/2025/1/e80140
doi:10.2196/80140

Peer Review of “COVID-19 Pneumonia Diagnosis Using Medical Images: Deep Learning-Based Transfer Learning Approach”

Ikenna Odezuligbo

Creighton University, Omaha, NE, United States

Related Articles:

Companion article: https://arxiv.org/abs/2503.12642v2

Companion article: https://med.jmirx.org/2025/1/e83230

Companion article: https://med.jmirx.org/2025/1/e75015

(JMIRx Med 2025;6:e83234) doi:10.2196/83234

KEYWORDS

computer vision; COVID-19 pneumonia diagnosis; deep learning; transfer learning; medical imaging analysis

This is the peer-review report for “COVID-19 Pneumonia Diagnosis Using Medical Images: Deep Learning-Based Transfer Learning Approach.”

Round 1 Review

General Comments

This manuscript [1] describes a transfer-learning approach using pretrained convolutional neural networks (VGG16, VGG19, ResNet-50) for binary COVID-19 detection on chest X-ray and computed tomography images. Overall, it tackles a timely problem and reports high accuracy (>97%), but several methodological and reporting issues limit confidence in the findings and their reproducibility.

Specific Comments

Major Comments

1. Lack of clinical validation: no in vivo or clinical ground-truth data are provided. The model’s >97% accuracy is based solely on public datasets; it’s unclear how it performs on real-world, heterogeneous clinical images.

2. Overfitting and hyperparameter tuning: identical performance across 5 hyperparameter settings for VGG16 suggests under- or overfitting. No learning curves or regularization impact analyses are shown to substantiate robustness claims.

3. Model comparison baseline: no comparison against simple baselines (eg, logistic regression on hand-crafted features) or recent literature benchmarks is provided, making it difficult to evaluate novelty and real gain.

Minor Comments

4. Repeated headings: “Integration into Mobile/Cloud-based Platform” appears twice in section 1; please consolidate.

5. Typographical and formatting errors: multiple sentences start without capitalization (eg, “we reviewing to the difference…”) and several references lack publication details (eg, [27,28] list only URLs).

Conflicts of Interest

None declared.

Reference

1. Dharmik A. COVID-19 pneumonia diagnosis using medical images: deep learning-based transfer learning approach. JMIRx Med 2025;6:e75015. [doi: 10.2196/75015]

Edited by F Wu; submitted 29.08.25; this is a non–peer-reviewed article;accepted 29.08.25; published 26.09.25.

Please cite as:
Odezuligbo I
Peer Review of "COVID-19 Pneumonia Diagnosis Using Medical Images: Deep Learning-Based Transfer Learning Approach"
JMIRx Med 2025;6:e83234
URL: https://xmed.jmir.org/2025/1/e83234
doi:10.2196/83234

Peer Review of “COVID-19 Pneumonia Diagnosis Using Medical Images: Deep Learning-Based Transfer Learning Approach”

Sunny Chi Lik Au

Tung Wah Eastern Hospital, So Kon Po, China (Hong Kong)

Related Articles:

Companion article: https://arxiv.org/abs/2503.12642v2

Companion article: https://med.jmirx.org/2025/1/e83230

Companion article: https://med.jmirx.org/2025/1/e75015

(JMIRx Med 2025;6:e83231) doi:10.2196/83231

KEYWORDS

computer vision; COVID-19 pneumonia diagnosis; deep learning; transfer learning; medical imaging analysis

This is the peer-review report for “COVID-19 Pneumonia Diagnosis Using Medical Images: Deep Learning-Based Transfer Learning Approach.”

Round 1 Review

General Comments

This paper [1] focused on the use of artificial intelligence (AI), in particular convolutional neural networks (CNNs) for detection of COVID-19 infections in radiological imaging. The study uses a substantial dataset of over 6000 images, which enhances the reliability of the results and supports robust model training and evaluation. Leveraging well-known CNNs such as VGG16, VGG19, and ResNet-50 demonstrates a practical application of transfer learning, a widely accepted technique in deep learning for medical imaging tasks.

However, in the Background and Introduction sections, the authors focused on the importance of rapid and early diagnosis of COVID-19, thus the demand for AI CNNs for diagnosis (“traditional diagnostic methods, such as serologic tests, have limitations, including low sensitivity and longer processing times”), yet this could be achieved easily nowadays with lateral flow devices or rapid antigen tests. Using computed tomography or X-rays to screen COVID-19 is far too expensive and time consuming compared to lateral flow devices or rapid antigen tests.

I believe the author was referring to the use of AI CNNs to differentiate COVID-19 pneumonia from other causes of pneumonia. Diagnosis of COVID-19 infection (which is usually mild and self-limiting) is totally different from COVID-19 pneumonia (which might require hospitalization and medical interventions). The authors might consider changing the title of the manuscript to “COVID 19 Pneumonia Diagnosis Analysis Using Transfer Learning–Deep Learning.” Similarly, for section 3.1, “COVID-19 Pneumonia Diagnosis Using Deep Learning” would be more appropriate than “COVID-19 Diagnosis Using Deep Learning.”

In addition, the Related Work section is brief and lacks depth. It does not sufficiently review existing medical studies on deep learning for COVID-19 pneumonia diagnosis, making it less comprehensive.

Specific Comments

Major Comments

1. Since this is a medical journal, medical terms are encouraged, for example, *anosmia* to replace *loss of smell*; *ageusia* to replace *loss of taste*.

2. Quite a significant number of references were not medical related, but related to AI or computer science. I would suggest the authors visit PubMed to search for more medical-related references. I cannot suggest any particular references to avoid conflicts of interest with certain groups of authors and to avoid self-citation.

3. The author detailed the AI CNN mechanism, yet the features of computed tomography or X-rays that were focused on were not mentioned. Was ground glass appearance the main target, or was it other features like cavitation, extent of lung involvement, or superior location? It would be more valid to evaluate various features targeted by the AI, instead of mentioning how it works.

Minor Comments

4. The author cited many different online references, yet the links or URLs were not available for readers to refer to. I would suggest adding the cited reference sources back for reviewers to assess the appropriateness of the citation, such as references 26 to 28, and for the benefit of readers. For example, reference 9 is not searchable on the internet.

5. In section 1.1, “At that point, there have been 98 confirmed cases and no reported deaths in 18 countries outside China...” Please add a reference citation for this factual statement.

6. In section 1.1, “As of 28 April 2020, 63% of worldwide mortality from the virus was from the Region...” Please clarify the “Region.”

7. In section 1.3, “Motivation to try to COVID-19 Diagnosis,” the English could be further polished, for example, “Motivation-to-try to COVID-19 Diagnosis” or “Motivation to try towards COVID-19 Diagnosis.”

8. *Computed tomography*, instead of *computer tomography*, is the proper term in section 3.1.

9. Abbreviations need not be spelled out again after their first use in the main text. For example, “computer tomography (CT)” in section 3.1 can just be “CT,” since CT has been defined already.

10. Please be consistent with reference citation formatting; various formats are used in the reference list, for example, “[20] Z. Wu et al Unsupervised feature learning via non-parametric instance discrimination. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 3733 - 3742, 2018”; “[24] Md. Islam, F. Karray, R. Alhajj, and J. Zeng. A review on deep learning techniques for the diagnosis of novel coronavirus (covid-19). IEEE Access, vol. 9, pp. 30551 - 30572, 2021. doi: 10.1109/ACCESS.2021.3058537”; and “[29] Mohammed K. Hassan, Ali I. El Desouky, Sally M. Elghamrawy, and Amany M. Sarhan. A hybrid real-time remote monitoring framework with nb-woa algorithm for patients with chronic diseases. https://doi.org/10.1016/j.future.2018.10.021, 2019. Future Generation Computer Systems, Volume 93, Pages 77 - 95, ISSN 0167 - 739X.”

11. To further improve the manuscript, please consider adding figures or tables showing the appearance of COVID-19 versus normal samples. Add to the Limitations section a discussion of potential biases (eg, dataset origin) or generalizability issues (eg, applicability to new variants) to demonstrate critical reflection

Round 2 Review

General Comments

Thanks for revising the manuscript according to the review comments. Most of my concerns are now addressed.

Specific Comments

Major Comments

1. Some parts of the manuscript used extensive bulleted lists; paragraphs should be used in the manuscript’s main text. If the author deems bullet points more appropriate for the content, the author could format lists as tables.

Conflicts of Interest

None declared.

Reference

1. Dharmik A. COVID-19 pneumonia diagnosis using medical images: deep learning-based transfer learning. JMIRx Med 2025;6:e75015. [doi: 10.2196/75015]

Abbreviations

AI: artificial intelligence
CNN: convolutional neural network

Edited by F Wu; submitted 29.08.25; this is a non–peer-reviewed article;accepted 29.08.25; published 26.09.25.

Please cite as:
Au SCL
Peer Review of “COVID-19 Pneumonia Diagnosis Using Medical Images: Deep Learning-Based Transfer Learning Approach”
JMIRx Med 2025;6:e83231
URL: https://xmed.jmir.org/2025/1/e83231
doi:10.2196/83231

Peer Review of “COVID-19 Pneumonia Diagnosis Using Medical Images: Deep Learning-Based Transfer Learning Approach”

Emmanuel Ndezure

Kwame Nkrumah University of Science and Technology, Kumasi, Ghana

Related Articles:

Companion article: https://arxiv.org/abs/2503.12642v2

Companion article: https://med.jmirx.org/2025/1/e83230

Companion article: https://med.jmirx.org/2025/1/e75015

(JMIRx Med 2025;6:e83236) doi:10.2196/83236

KEYWORDS

computer vision; COVID-19 pneumonia diagnosis; deep learning; transfer learning; medical imaging analysis

This is the peer-review report for “COVID-19 Pneumonia Diagnosis Using Medical Images: Deep Learning-Based Transfer Learning Approach.”

Round 1 Review

General Comments

This manuscript [1] investigates the application of deep learning, particularly transfer learning using VGG16, VGG19, and ResNet-50, for diagnosing COVID-19 through computed tomography and and X-ray images. The topic is important and timely, especially considering the enduring threat of COVID-19 variants and the burden on global health care systems. The author demonstrates technical familiarity with deep learning techniques, model tuning, and performance evaluation. However, there are areas where the study could be improved to enhance its rigor, clarity, and impact.

Specific Comments

Major Comments

1. Dataset description and bias: the paper mentions using a dataset of 6259 images (4651 COVID-19 cases and 1608 normal cases). However, there is no discussion on potential biases in the dataset, such as the source of the images, demographic diversity (age, gender, and geographic location), or the balance between COVID-19 and normal cases. Addressing these aspects would strengthen the validity of the results. I suggest that the author include a detailed description of the dataset, including sources, demographic information, and steps taken to mitigate bias, and consider discussing the imbalance in the dataset and how it might affect model performance.

2. Comparative analysis with existing methods: while the paper reports high accuracy (97.73%) for the proposed models, it lacks a comprehensive comparison with other state-of-the-art methods or baseline models. This makes it difficult to assess the novelty and superiority of the proposed approach. I suggest that the author add a comparative table or section that contrasts the performance of VGG16, VGG19, and ResNet-50 with other recent studies or baseline models and highlight the unique contributions of this work.

3. Clinical relevance and practical deployment: the study focuses on technical performance metrics but does not discuss the clinical applicability of the models. For instance, how would these models integrate into real-world health care settings? What are the potential challenges (eg, computational resources, interpretability for clinicians)? I suggest that the author expand the discussion on clinical relevance, including limitations and practical considerations for deployment in health care systems.

4. Language and grammar: the manuscript needs extensive language editing. There are frequent grammatical issues, awkward phrasing (eg, “the 1608 belong to healthy people”), and repetition. A professional edit is highly recommended to improve readability and flow.

5. Figures and tables: several figures (eg, confusion matrices, loss/accuracy curves) are referenced but lack sufficient clarity, labeling, or captions. Figures 4 to 8 must be embedded clearly within the results discussion and interpreted to guide the reader. Ensure figures are high resolution and correctly formatted.

6. Overstatement of results: the paper claims high performance (97.73% accuracy), yet offers little discussion on external validity or overfitting risks. Since cross-validation was performed on a relatively small dataset, these results may not generalize well. The author should tone down claims and discuss limitations.

7. Dataset description and ethics: while the dataset is described as publicly available, the manuscript lacks ethical approval or justification. Clarify whether ethical clearance was required. Also, organize the dataset description into a single, detailed section including data sources, balance between classes, preprocessing applied, and augmentation steps.

8. Evaluation metrics and statistical rigor: the paper heavily relies on accuracy, sensitivity, specificity, and F_1-score, but fails to report CIs or conduct statistical tests to validate performance differences between models. Including receiver operating characteristic area under the curve values and visualizations would also strengthen the evaluation.

9. Novelty and contribution not clearly established: while the paper uses popular convolutional neural network architectures, there is no clear indication of what is novel in this study compared to the extensive body of existing work using these same models on similar datasets. What distinguishes this work? Is it the dataset size, preprocessing technique, tuning strategy, or model ensemble?

Minor Comments

10. Hyperparameter tuning details: the paper describes hyperparameter tuning but does not explain the rationale behind the selected ranges (eg, learning rate and batch size). A brief justification for these choices would improve reproducibility. I suggest adding a sentence or two explaining why the specified ranges for hyperparameters were chosen.

11. Use consistent terminology throughout (eg, “deep learning model” versus “CNN-based model”).

12. Data augmentation techniques: these are described generically. Specify which augmentations were applied and how frequently. Were augmentation parameters validated?

13. Please structure the abstract under clear headings, Background, Objective, Methods, Results, and Conclusion, to aid clear reading and comprehension.

Round 2 Review

Specific Comments

Major Comments

I commend the author for the comprehensive revisions made in response to the initial review. The manuscript now demonstrates significant improvements in clarity, organization, and scientific rigor. Key concerns, including dataset bias, comparative evaluation with existing methods, clinical applicability, language quality, and statistical robustness, have been adequately addressed. Figures and tables, which were initially submitted as separate files, have now been properly embedded and contextualized within the main manuscript, greatly improving readability and interpretation of results.

All my previous comments have been satisfactorily responded to, and I have no further critical concerns. I find the revised manuscript suitable for publication.

Conflicts of Interest

None declared.

Reference

1. Dharmik A. COVID-19 pneumonia diagnosis using medical images: deep learning-based transfer learning approach. JMIRx Med 2025;6:e75015. [doi: 10.2196/75015]

Edited by F Wu; submitted 29.08.25; this is a non–peer-reviewed article;accepted 29.08.25; published 26.09.25.

Please cite as:
Ndezure E
Peer Review of “COVID-19 Pneumonia Diagnosis Using Medical Images: Deep Learning-Based Transfer Learning Approach”
JMIRx Med 2025;6:e83236
URL: https://xmed.jmir.org/2025/1/e83236
doi:10.2196/83236

Peer-Review of "Real-Time Health Monitoring Using 5G Networks: Deep Learning–Based Architecture for Remote Patient Care"

Shruti Bharadwaj

United College of Engineering & Research, United Tower 53, Leader Road, Prayagraj, India

Related Articles:

Companion article: https://arxiv.org/abs/2501.01027v1

Companion article: https://med.jmirx.org/2025/1/e83473

Companion article: https://med.jmirx.org/2025/1/e70906

(JMIRx Med 2025;6:e83423) doi:10.2196/83423

KEYWORDS

5G; real-time patient monitoring; vital signs; prediction; deep learning; machine learning

This is a peer-review report for "Real-Time Health Monitoring Using 5G Networks: Deep Learning–Based Architecture for Remote Patient Care."

Round 1 Review

Review Report

This paper [1] presents a novel architecture integrating deep learning and 5G networks to enhance real-time remote patient monitoring.

1. The combination of a convolutional neural network/long short-term memory model with 5G ultra-reliable low latency communication enables real-time monitoring with high accuracy and low latency. Achieving 96.5% accuracy for vital sign prediction demonstrates the effectiveness of the proposed model.
2. While tested on 1000 patients, analysis of its scalability to larger populations with diverse demographics would improve generalizability.
3. The use of attention mechanisms in the long short-term memory component improves the system's ability to model dependencies in continuous vital sign monitoring.
4. A more detailed comparison with state-of-the-art remote monitoring systems, including their architectures and limitations, would strengthen the claims.
5. Since patient data is transmitted over 5G networks, an evaluation of encryption techniques, data integrity measures, and compliance with health care regulations (eg, the Health Insurance Portability and Accountability Act and the General Data Protection Regulation) should be included. Investigating performance under network congestion, packet loss, or fluctuations in 5G coverage would ensure system reliability.

Final Recommendation

Accept with minor revisions.

This paper presents a promising and well-structured approach to real-time patient monitoring using deep learning and 5G technology. However, addressing concerns regarding computational efficiency, scalability, security, robustness, and explainability would further strengthen its impact.

Suggested Revisions

- Include a comparative analysis with other remote patient monitoring systems.
- Provide details on computational resource use and energy efficiency for edge deployment.
- Address security, encryption, and data privacy considerations.
- Test and discuss model performance under varying network conditions.

Conflicts of Interest

None declared.

Reference

1. Batool I. Real-time health monitoring using 5G neworks: deep learning–based architecture for remote patient care. JMIRx Med 2025;6:e70906. [doi: 10.2196/70906]

Edited by A Grover; submitted 02.09.25; this is a non–peer-reviewed article;accepted 02.09.25; published 01.10.25.

Please cite as:
Bharadwaj S
Peer-Review of "Real-Time Health Monitoring Using 5G Networks: Deep Learning–Based Architecture for Remote Patient Care"
JMIRx Med 2025;6:e83423
URL: https://xmed.jmir.org/2025/1/e83423
doi:10.2196/83423

Peer Review of "Real-Time Health Monitoring Using 5G Networks: Deep Learning–Based Architecture for Remote Patient Care"

Francisco Javier Gonzalez-Canete

Universidad de Malaga, Avda. Cervantes, Malaga, Spain

Related Articles:

Companion article: https://arxiv.org/abs/2501.01027v1

Companion article: https://med.jmirx.org/2025/1/e83473

Companion article: https://med.jmirx.org/2025/1/e70906

(JMIRx Med 2025;6:e83424) doi:10.2196/83424

KEYWORDS

5G; real-time patient monitoring; vital signs; prediction; deep learning; machine learning

This is a peer-review report for "Real-Time Health Monitoring Using 5G Networks: Deep Learning–Based Architecture for Remote Patient Care."

Round 1 Review

General Comments

This paper [1] proposes an architecture that integrates a deep learning method with 5G networks to monitor health parameters. These parameters are analyzed using a deep learning system located in the edge network. The decisions taken from the edge system are transfered to a medical server using the 5G network. The objective is to obtain a high-accuracy evaluation of the deep learning system and obtain very low latency in data transmission and reception.

Specific Comments

Major Comments

1. Table 3 is not referenced nor commented on in the text. You should add a paragraph explaining the table or delete it.
2. Table 5 compares the system performance with 3 other systems, A, B, and C, but those systems are never described. They must be commented on in order to compare results.

Minor Comments

1. Equation 1 has no label (1) and it is defined twice.
2. Figure 4 should be placed after it is called out.
3. On page 6, there is a sentence in square brackets.
4. Correct the sentence "Figure 4illustrates..." The number 4 and the word "illustrates" are too close.
5. Table 5 is called out before Table 4. Consequently, they should be switched.
6. The sentence "Table V System Comparison..." seems to be a figure description instead of part of the text. It makes no sense in the place it is located.
7. The text "(P ! .001)" I presume should be "(P<.001)"

Round 2 Review

General Comments

This paper presents an architecture to perform real-time monitoring of health signals using 5G networks and deep-learning prediction of possible health problems using the aquired signals.

Specific Comments

Major Comments

1. There are some equations with no defined parameters. In equation 16, what are *Pij* and *xij*? In equation 17, what is *N*? In equation 18, what are *Bi*, *Cj*, and *M*? In equation 19, what is *Lu*? They must be defined.
2. How are weights *wu*, *wr*, and *wl* calculated or estimated? What are their chosen values? The final performance could change depending on the selection of these parameters, as you are giving more importance to one parameter or another.

Minor Comments

1. Most of the references are "touching" the previous text. Add a space between text and references. For instance: "...clinical settings[1],[2]." should be "... clinical settings [1], [2]."
2. Figure 2 should be closer to where it is referred to on the previous page.

Conflicts of Interest

None declared.

Reference

1. Batool I. Real-time health monitoring using 5G neworks: deep learning–based architecture for remote patient care. JMIRx Med 2025;6:e70906. [doi: 10.2196/70906]

Edited by A Grover; submitted 02.09.25; this is a non–peer-reviewed article;accepted 02.09.25; published 01.10.25.

Please cite as:
Gonzalez-Canete FJ
Peer Review of "Real-Time Health Monitoring Using 5G Networks: Deep Learning–Based Architecture for Remote Patient Care"
JMIRx Med 2025;6:e83424
URL: https://xmed.jmir.org/2025/1/e83424
doi:10.2196/83424

Peer Review of “Real-World Performance of COVID-19 Antigen Tests: Predictive Modeling and Laboratory-Based Validation”

Gerald Kost
UC Davis, Davis, CA, United States

Related Articles:

Companion article: https://www.medrxiv.org/content/10.1101/2024.10.21.24315762v1

Companion article: https://med.jmirx.org/2025/1/e83474

Companion article: https://med.jmirx.org/2025/1/e68376

(JMIRx Med 2025;6:e83479) doi:10.2196/83479

KEYWORDS

COVID-19; real-world data; limit of detection lateral flow test; probability of positive agreement; logistic regression; COVID-19 antigen test clinical performance

This is a peer-review report for “Real-World Performance of COVID-19 Antigen Tests: Predictive Modeling and Laboratory-Based Validation.”

Round 1 Review

General Comments

This paper [1] introduces a novel approach to speeding COVID-19 antigen test deployment, possibly during the next pandemic, whether it be a COVID-19 variant or a new pathogen that arises.

Specific Comments

Major Comments

1. The authors’ clinical conclusions based on their prediction theory are overly optimistic.

2. The authors can explore actual clinical evaluations to determine the robustness of their prediction modeling.

3. Thus, the paper merits publication, providing the limitations are more clearly described and the conclusions are limited to the mathematical results for which the authors have proven their claims theoretically. Extension to clinical applicability is a different story yet to be told.

4. The authors should be encouraged to move forward in view of the need and the poor performance of COVID-19 rapid antigen tests during the pandemic because of low sensitivity, a lack of deep understanding, and the “prevalence boundary,” a measure of when the rate of false omissions becomes too high and false negatives spread disease.

Minor Comments

5. Needs English grammar review. This could be achieved by using an artificial intelligence editor.

Conflicts of Interest

None declared.

Reference

1. Bosch M, Garcia D, Rudtner L, et al. Real-world performance of COVID-19 antigen tests: predictive modeling and laboratory-based validation. JMIRx Med 2025;6:e68376. [doi: 10.2196/68376]

Edited by F Wu; submitted 03.09.25; this is a non–peer-reviewed article;accepted 03.09.25; published 06.10.25.

Please cite as:
Kost G
Peer Review of “Real-World Performance of COVID-19 Antigen Tests: Predictive Modeling and Laboratory-Based Validation”
JMIRx Med 2025;6:e83479
URL: https://xmed.jmir.org/2025/1/e83479
doi:10.2196/83479

Peer Review of “Real-World Performance of COVID-19 Antigen Tests: Predictive Modeling and Laboratory-Based Validation”

Helena de Puig

Wyss Institute for Biologically Inspired Engineering, Harvard University, Boston, MA, United States

Related Articles:

Companion article: https://www.medrxiv.org/content/10.1101/2024.10.21.24315762v1

Companion article: https://med.jmirx.org/2025/1/e83474

Companion article: https://med.jmirx.org/2025/1/e68376

(JMIRx Med 2025;6:e83476) doi:10.2196/83476

KEYWORDS

COVID-19; real-world data; limit of detection lateral flow test; probability of positive agreement; logistic regression; COVID-19 antigen test clinical performance

This is a peer-review report for “Real-World Performance of COVID-19 Antigen Tests: Predictive Modeling and Laboratory-Based Validation.”

Round 1 Review

This is a good paper [1] that models the performance of antigen tests.

Minor Comments

1. It would be good to include a schematic/analysis/methods and an image of how the lateral flow assays look and how test band intensities are obtained. Was that done with ImageJ?
2. An interesting aspect of the paper is the comparison between trained eye versus user of lateral flow assays (Figure 5). I think that adding a paragraph about the conclusions from that figure might be good in the Discussion section.

Conflicts of Interest

None declared.

Reference

1. Bosch M, Garcia D, Rudtner L, et al. Real-world performance of COVID-19 antigen tests: predictive modeling and laboratory-based validation. JMIRx Med 2025;6:e68376. [doi: 10.2196/68376]

Edited by F Wu; submitted 03.09.25; this is a non–peer-reviewed article;accepted 03.09.25; published 06.10.25.

Please cite as:
de Puig H
Peer Review of “Real-World Performance of COVID-19 Antigen Tests: Predictive Modeling and Laboratory-Based Validation”
JMIRx Med 2025;6:e83476
URL: https://xmed.jmir.org/2025/1/e83476
doi:10.2196/83476

Peer Review of "Effects of Interventions for the Prevention and Management of Maternal Anemia in the Advent of the COVID-19 Pandemic: Systematic Review and Meta-Analysis"

Suriya Kumareswaran

Johor State Health Department, Johor, Malaysia

Related Articles:

Companion article: https://www.medrxiv.org/content/10.1101/2024.02.18.24302492v1

Companion article: https://med.jmirx.org/2025/1/e81711

Companion article: https://med.jmirx.org/2025/1/e57626

(JMIRx Med 2025;6:e81699) doi:10.2196/81699

KEYWORDS

maternal anemia; anemia in pregnancy; COVID-19; pregnancy complications; meta-analysis; maternal and child health; anemia prevention; reproductive health

This is the peer-review report for "Effects of Interventions for the Prevention and Management of Maternal Anemia in the Advent of the COVID-19 Pandemic: Systematic Review and Meta-Analysis."

Round 1 Review

General Impressions

The study [1] addresses a crucial public health issue—maternal anemia—and its dynamics in the context of the COVID-19 pandemic. The topic is relevant and timely, particularly given the pandemic's disruptive effect on health care systems worldwide. However, the manuscript has several shortcomings that require significant revisions for clarity, coherence, and scientific rigor. Below, I provide a detailed assessment with major comments and minor comments for improvement.

Major Comments

1. Scientific rigor and novelty.

Strength: The focus on maternal anemia interventions during COVID-19 is unique and addresses a significant gap in the literature.

Issue: The study does not establish the novelty of its findings clearly. It cites several similar meta-analyses but does not differentiate its contribution.

Recommendation: Clarify how this meta-analysis advances existing knowledge. Are there new methodologies, expanded datasets, or novel insights?

2. Study design and methodology.

Inclusion criteria: The inclusion of preprints and unpublished data raises concerns about the reliability and quality of the evidence.

Suggestion: Clearly discuss the rationale for including preprints and outline strategies to mitigate biases.

Subgroup analysis: While subgroup analyses are insightful, the interpretation of heterogeneity (I^2>90% in multiple cases) is not adequately addressed. The sensitivity analyses seem to mitigate this but are not discussed in sufficient depth.

Suggestion: Incorporate a robust discussion of the potential sources of heterogeneity and its implications for the results.

3. Data presentation.

Tables and figures: Tables and figures are overly complex and lack clarity.

Suggestion: Simplify forest and funnel plots for better readability. Ensure that all figures are annotated clearly.

Forest plots: Some rate ratio confidence intervals (eg, in subgroup analysis) overlap with no-effect lines, which undermines conclusions about statistical significance.

Suggestion: Address these overlaps explicitly in the Discussion.

4. Statistical analysis.

Publication bias: The funnel plots indicate substantial publication bias. This is acknowledged but inadequately addressed in the Discussion.

Suggestion: Include a deeper discussion of how this bias impacts the reliability of pooled estimates.

Fixed- versus random-effects models: The rationale for choosing fixed- or random-effects models for different analyses is not well-articulated.

Suggestion: Explain this choice clearly, especially in the context of high heterogeneity.

5. Interpretation of results.

The interpretation of intervention effects (eg, a 17% improvement for iron supplementation) does not account for clinical significance, which may differ from statistical significance.

Suggestion: Discuss the practical implications of these findings, especially in low-resource settings.

6. Language and readability.

The manuscript is riddled with grammatical errors, unclear phrasing, and redundancies. For instance:

"The effect on prevention, control, management and or treatment of anemia was calculated and compared between the intervention and the comparator arms."

Suggestion: Simplify and clarify language to improve readability.

Acronyms (eg, RR, CI, IFA) are used without clear explanation.

Suggestion: Ensure all acronyms are defined upon first use.

7. Ethical considerations.

The manuscript mentions that some data are unpublished. It is unclear whether these studies adhered to ethical guidelines.

Suggestion: Add a section on ethical considerations, particularly around the inclusion of unpublished studies.

8. Discussion and Conclusion.

Weakness: The Discussion is repetitive and does not critically engage with the limitations of the study or the broader implications of the findings.

Suggestion: Provide a more focused discussion of limitations (eg, high heterogeneity, reliance on observational studies), implications for practice and policy, and recommendations for future research.

Minor Comments

1. Abstract.

Issue: The abstract lacks precision and overuses vague terms (eg, "several anemia interventions").

Suggestion: Summarize key findings clearly, avoiding overgeneralizations.

2. Introduction.

The Introduction is overly lengthy and includes redundant information (eg, definitions of anemia repeated multiple times).

Suggestion: Streamline the Introduction to focus on the problem, the gap in knowledge, and the study's objectives.

3. References.

References are incomplete and inconsistently formatted.

Suggestion: Ensure all references follow a standardized format (eg, APA, AMA).

4. Figures.

Figures are not numbered or titled appropriately.

Suggestion: Include clear figure numbers, titles, and legends for all figures.

Recommendation for Authors

Based on the above assessment, this manuscript requires major revisions. Key issues include addressing heterogeneity and publication bias in statistical analysis, improving clarity and rigor in data presentation, and enhancing language and readability.

Conflicts of Interest

None declared.

Reference

1. Muthuka JK, Mbari-Fondo DK, Wambura FM, et al. Effects of interventions for the prevention and management of maternal anemia in the advent of the COVID-19 pandemic: systematic review and meta-analysis. JMIRx Med 2025;6:e57626. [doi: 10.2196/57626]

Edited by E Meinert, T Leung; submitted 01.08.25; this is a non–peer-reviewed article;accepted 01.08.25; published 06.10.25.

Please cite as:
Kumareswaran S
Peer Review of "Effects of Interventions for the Prevention and Management of Maternal Anemia in the Advent of the COVID-19 Pandemic: Systematic Review and Meta-Analysis"
JMIRx Med 2025;6:e81699
URL: https://xmed.jmir.org/2025/1/e81699
doi:10.2196/81699

Peer Review of “Effects of Interventions for the Prevention and Management of Maternal Anemia in the Advent of the COVID-19 Pandemic: Systematic Review and Meta-Analysis”

Anonymous

Related Articles:

Companion article: https://www.medrxiv.org/content/10.1101/2024.02.18.24302492v

Companion article: https://med.jmirx.org/2025/1/e81711

Companion article: https://med.jmirx.org/2025/1/e57626

(JMIRx Med 2025;6:e82836) doi:10.2196/82836

KEYWORDS

maternal anemia; anemia in pregnancy; COVID-19; pregnancy complications; meta-analysis; maternal and child health; anemia prevention; reproductive health

This is the peer-review report for “Effects of Interventions for the Prevention and Management of Maternal Anemia in the Advent of the COVID-19 Pandemic: Systematic Review and Meta-Analysis.”

Round 1 Review

Some weaknesses, gaps, and limitations in this study [1] based on the current study provided could include:

1. Retrospective studies: The majority of the studies included in the meta-analysis were retrospective epidemiological studies, which may have limitations in terms of bias, data accuracy, and causality compared to prospective studies or randomized controlled trials. This could affect the reliability and generalizability of the findings.
2. Heterogeneity: The high heterogeneity identified in the pooled effect estimates suggests variability in study designs, interventions, and outcomes across the included studies. This heterogeneity can impact the interpretation of the results and the ability to draw consistent conclusions.
3. Publication bias: The presence of publication bias indicated by the asymmetrical funnel plot could introduce bias in the pooled effect estimates. This bias may be due to the selective reporting of studies with significant results, potentially skewing the overall findings.
4. Limited scope: Some studies may not have clearly defined the age range of participants or the specific stage of the gestation period analyzed. A lack of detailed information on these aspects could limit the applicability and generalizability of the results to specific subgroups of pregnant women.
5. Indirect effects of COVID-19: While the study focused on the direct impact of COVID-19 on maternal anemia interventions, indirect contributions of the pandemic on anemic conditions may not have been fully elucidated. Understanding these indirect effects could provide a more comprehensive view of the challenges faced during the pandemic.
6. Effectiveness trends: The decreasing trend in the effectiveness of interventions against maternal anemia from 2020 to 2022 raises questions about the sustainability and adaptability of intervention strategies, especially in the context of global health emergencies. Further research is needed to explore the reasons behind this trend and potential strategies for improvement.

Addressing these weaknesses and limitations could enhance the validity and applicability of the study findings and contribute to a more comprehensive understanding of the impact of COVID-19 on maternal anemia interventions.

Overall, the meta-analysis highlighted the challenges faced in addressing maternal anemia during the COVID-19 pandemic and the need for continuous monitoring and adaptation of intervention strategies to mitigate adverse outcomes.

Conflicts of Interest

None declared.

Reference

1. Muthuka JK, Mbari-Fondo DK, Wambura FM, et al. Effects of interventions for the prevention and management of maternal anemia in the advent of the COVID-19 pandemic: systematic review and meta-analysis. JMIRx Med 2025;6:e57626. [doi: 10.2196/57626]

Edited by E Meinert, T Leung; submitted 22.08.25; this is a non–peer-reviewed article;accepted 22.08.25; published 06.10.25.

Please cite as:
Anonymous
Peer Review of "Effects of Interventions for the Prevention and Management of Maternal Anemia in the Advent of the COVID-19 Pandemic: Systematic Review and Meta-Analysis"
JMIRx Med 2025;6:e82836
URL: https://xmed.jmir.org/2025/1/e82836
doi:10.2196/82836

Peer Review of “Effects of Interventions for the Prevention and Management of Maternal Anemia in the Advent of the COVID-19 Pandemic: Systematic Review and Meta-Analysis”

I Gde Sastra Winata

Udayana University, Bali, Indonesia

Related Articles:

Companion article: https://www.medrxiv.org/content/10.1101/2024.02.18.24302492v1

Companion article: https://med.jmirx.org/2025/1/e81711

Companion article: https://med.jmirx.org/2025/1/e57626

(JMIRx Med 2025;6:e81700) doi:10.2196/81700

KEYWORDS

maternal anemia; anemia in pregnancy; COVID-19; pregnancy complications; meta-analysis; maternal and child health; anemia prevention; reproductive health

This is the peer-review report for “Effects of Interventions for the Prevention and Management of Maternal Anemia in the Advent of the COVID-19 Pandemic: Systematic Review and Meta-Analysis.”

Round 1 Review

General Comments

This paper [1] addresses a relevant and important topic in maternal health, particularly in the context of anemia during pregnancy. The paper has potential for publication in a journal with minor revisions.

Specific Comments

Major Comments

- 1. There is no conclusion in this manuscript. Add a Conclusion section that summarizes the content of this study.

Minor Comments

- 2. Introduction section. Briefly explain the types of interventions implemented during COVID-19 to prevent anemia in pregnant women. Provide a brief explanation of the differences in anemia prevention interventions before and after COVID-19.
- 3. Discussion, at the end of the Discussion section. Include the main findings of this study and emphasize their significance in addressing anemia-related challenges. Highlight the contribution of the study results to public health, particularly how the findings can inform or improve anemia prevention and treatment strategies in health care systems. Provide practical recommendations or actionable steps based on the study’s outcomes that can be implemented in maternal health care policies and programs.

Conflicts of Interest

None declared.

Reference

1. Muthuka JK, Mbari-Fondo DK, Wambura FM, et al. Effects of interventions for the prevention and management of maternal anemia in the advent of the COVID-19 pandemic: systematic review and meta-analysis. JMIRx Med 2025;6:e57626. [doi: 10.2196/57626]

Edited by E Meinert, T Leung; submitted 01.08.25; this is a non–peer-reviewed article;accepted 01.08.25; published 06.10.25.

Please cite as:
Winata IGS
Peer Review of "Effects of Interventions for the Prevention and Management of Maternal Anemia in the Advent of the COVID-19 Pandemic: Systematic Review and Meta-Analysis"
JMIRx Med 2025;6:e81700
URL: https://xmed.jmir.org/2025/1/e81700
doi:10.2196/81700

Peer Review of “Estimating Variance of Log Standardized Incidence Ratios Assessing Health Care Providers’ Performance: Comparative Analysis Using Bayesian, Bootstrap, and Delta Method Approaches”

Emmanuel Oluwagbade

Vanderbilt University, Nashville, TN, United States

Related Articles:

Companion article: https://www.medrxiv.org/content/10.1101/2025.04.22.25326183v1

Companion article: https://med.jmirx.org/2025/1/e83796

Companion article: https://med.jmirx.org/2025/1/e77415

(JMIRx Med 2025;6:e83798) doi:10.2196/83798

KEYWORDS

standardized incidence ratio; SIR; performance; health care providers; machine learning; equity

This is a peer-review report for “Estimating Variance of Log Standardized Incidence Ratios Assessing Health Care Providers’ Performance: Comparative Analysis Using Bayesian, Bootstrap, and Delta Method Approaches.”

Round 1 Review

General Comments

This paper [1] compares 3 approaches for estimating the variance of the log standardized incidence ratio when profiling kidney replacement therapy centers in Australia: (1) the analytical delta method, (2) the nonparametric bootstrap method (5000 resamples), and (3) Bayesian Markov chain Monte Carlo (25,500 iterations, 3 chains). Using 2005 - 2023 patient-level data from the Australia and New Zealand Dialysis and Transplant Registry and a random-effects logistic model, the authors evaluated bias, variance, and mean squared error (MSE) and visualized performance via funnel plots. Results indicated similar bias across methods but substantially lower variance and MSE for the Markov chain Monte Carlo method (bias≈0.019; variance=0.00005; MSE=0.00042) compared with the bootstrap method (variance=0.00027; MSE=0.00094).

The topic is practical and timely, yet the manuscript needs clearer model specification, interval coverage evaluation, and streamlined writing before it reaches publishable quality.

Specific Comments

Major Comments

1. There are some concerns around model clarity (Poisson versus logistic language mixed; appendix is missing). Provide complete model equations, a covariate list, and software/code links and justify using the Bernoulli model for a ratio outcome.
2. Interval coverage and type I error absent. Action: add a simulation or internal bootstrap to report 95% interval coverage and false-positive rates for each method.
3. Missing data handling unexplained.Action: quantify missingness, describe any imputation, and list all risk-adjustment covariates.

Minor Comments

In Table 1, add units and align decimals.

Conflicts of Interest

None declared.

Reference

1. Woldeyohannes S, Jones Y, Lawton P. Estimating variance of log standardized incidence ratios assessing health care providers’ performance: comparative analysis using Bayesian, bootstrap, and delta method approaches. JMIRx Med 2025;6:e77415. [doi: 10.2196/77415]

Edited by S Tungjitviboonkun; submitted 08.09.25; this is a non–peer-reviewed article;accepted 08.09.25; published 09.10.25.

Please cite as:
Oluwagbade E
Peer Review of "Estimating Variance of Log Standardized Incidence Ratios Assessing Health Care Providers' Performance: Comparative Analysis Using Bayesian, Bootstrap, and Delta Method Approaches"
JMIRx Med 2025;6:e83798
URL: https://xmed.jmir.org/2025/1/e83798
doi:10.2196/83798

Peer Review of “Development of a Conversational Artificial Intelligence–Based Web Application for Medical Consultations: Prototype Study”

Anonymous

Related Articles:

Companion article: https://www.medrxiv.org/content/10.1101/2023.12.31.23300681v1

Companion article: https://med.jmirx.org/2025/1/e83417

Companion article: https://med.jmirx.org/2025/1/e56090

(JMIRx Med 2025;6:e83217) doi:10.2196/83217

KEYWORDS

artificial intelligence; ChatGPT; chatbots; conversational agent; machine learning

This is the peer-review report for “Development of a Conversational Artificial Intelligence–Based Web Application for Medical Consultations: Prototype Study.”

Round 1 Review

Specific Comments

Major Comments

1. I was really liking the idea of this paper [1] and read it with great interest, but perhaps I misunderstood—I was hoping it was a chatbot that would actually give me a diagnosis (eg, once I input an image of my retina or have some conversation with it) rather than just referring me to a specialist (which would be appropriate in some cases). Please correct my understanding if I am wrong.
2. From the initial paper idea, I got the feeling that it was going to be an app where I could start uploading images of X-rays and the chatbot, using its models, would start to tell me something about the image; instead, it seems from the example figures that all it is doing is telling the user that this is an X-ray image and to contact an expert. I am not sure how this is even useful? Perhaps I read the paper in haste and am lacking understanding. I would suggest showcasing a full conversation from each of your areas (X-rays, diabetes, etc), with a full screen capture of the conversation, showing an image uploaded and ending with a diagnosis (if indeed that is what your bot is capable of).
3. Figures 2-5 do not really give me any picture of what is going on, they just reaffirm what I thought; that is, that the bot is not actually giving any information except recognizing what type of image it is, then referring the user to a consultant? Is that correct? You really need to put some nice figures of your full flow and architecture, not too complex, but the ones you show do not really, in my opinion, provide the reader with any real value. As a reader, I want to see what it is you have done, and as a technical person who wants to replicate your work, I would want to see your architecture in diagram form, as well as a proper flowchart of some sort (again, no need to be complex, but to a high standard as you normally see in leading journals) outlining exactly what the flow is; this should correspond to the actual app screen captures so readers can see exactly what your app does.
4. Overall summary: Having worked on and researched chatbots, I read this with great interest but, as per my comments, I am a little confused, as it seems this bot simply refers the user to a person after recognizing an image as an X-ray, for example. I was under the impression from the content or was half expecting the ability to input an image, be that of an X-ray or retina, and it would start giving me some diagnostic information or the like.

Minor Comments

1. The manuscript is written in some places more like a personal blog; this should be changed and be more in line with how journal articles are written.
2. Many grammatical errors and some spelling mistypes, such as writing “chess” instead of “chest”; generally, it needs to be professionally proofread and the language tidied.

Round 2 Review

General Comments

The author has given a response to each point and I am satisfied.

Conflicts of Interest

None declared.

Reference

1. Pires JG. Development of a Conversational Artificial Intelligence–Based Web Application for Medical Consultations: Prototype Study. JMIRx Med 2025;URL:e56090. [doi: 10.2196/56090]

Edited by T Leung; submitted 29.08.25; this is a non–peer-reviewed article;accepted 29.08.25; published 15.10.25.

Please cite as:
Anonymous
Peer Review of "Development of a Conversational Artificial Intelligence–Based Web Application for Medical Consultations: Prototype Study"
JMIRx Med 2025;6:e83217
URL: https://xmed.jmir.org/2025/1/e83217
doi:10.2196/83217

Peer Review of “Development of a Conversational Artificial Intelligence–Based Web Application for Medical Consultations: Prototype Study”

Anonymous

Related Articles:

Companion article: https://www.medrxiv.org/content/10.1101/2023.12.31.23300681v1

Companion article: https://med.jmirx.org/2025/1/e83417

Companion article: https://med.jmirx.org/2025/1/e56090

(JMIRx Med 2025;6:e84443) doi:10.2196/84443

KEYWORDS

artificial intelligence; ChatGPT; chatbots; conversational agent; machine learning

This is the peer-review report for “Development of a Conversational Artificial Intelligence–Based Web Application for Medical Consultations: Prototype Study.”

Round 1 Review

General Comments

This paper [1] presents a novel conversational artificial intelligence (AI) that is built on top of several models capable of detecting various health conditions. The research itself is interesting, relevant, and carried out well. I look forward to seeing the revised work!

Specific Comments

Major Comments

1. Please strongly consider rechecking the grammar for the paper as I found several mistakes and inconsistencies in just the Abstract alone. Grammar issues have made it difficult to follow several lines of thought throughout the paper and have lowered its overall quality. Furthermore, the style of writing is more conversational than it is formal and academic in several cases. This is my main reason for rejecting the paper. Please address these issues and then resubmit.
2. Because the experience is integrated into a chatbot, indicating that providing an interactive user experience was a goal of this work, it would be good to also conduct some user research to assess the usability of the system and participants’ impressions of it.
3. It is important to also include a section discussing the potential dangers and ethical implications of deploying such a chatbot in the real world given the sensitive context and its critical implications.

Minor Comments

1. I find the opening sentence of the Abstract and Introduction (“Artificial intelligence (AI) evolved in trends. Currently, the trend is conversational artificial intelligence (CAI).”) to be problematic. It is unclear what the statement “AI evolved in trends” means and therefore it is difficult to evaluate its accuracy. Furthermore, it would be more apt to say that the trend now is generative AI, which translates to large language models, that is fuelling conversational AI. It is also worth noting that conversational AI is not just focused on text-based interactions, but also includes voice modalities.

Round 2 Review

General Comments

Thank you to the authors for reading our comments and revising the paper. Many of our previous comments have been addressed; however, I still believe the writing style, grammar, and language of the paper need significant work before this can be published.

Conflicts of Interest

None declared.

Reference

1. Pires JG. Development of a Conversational Artificial Intelligence–Based Web Application for Medical Consultations: Prototype Study. JMIRx Med 2025;URL:e56090. [doi: 10.2196/56090]

Abbreviations

AI: artificial intelligence

Edited by T Leung; submitted 19.09.25; this is a non–peer-reviewed article;accepted 19.09.25; published 15.10.25.

Please cite as:
Anonymous
Peer Review of "Development of a Conversational Artificial Intelligence–Based Web Application for Medical Consultations: Prototype Study"
JMIRx Med 2025;6:e84443
URL: https://xmed.jmir.org/2025/1/e84443
doi:10.2196/84443

Peer Review of “Impact of the COVID-19 Pandemic on Routine Childhood Vaccination Coverage in Ecuador From 2019 to 2021: Comparative Analysis”

Adeleke Adekola

Syracuse University, Syracuse, NY, United States

Related Articles:

Companion article: https://www.medrxiv.org/content/10.1101/2025.03.26.25324742v1

Companion article: https://med.jmirx.org/2025/1/e84851

Companion article: https://med.jmirx.org/2025/1/e75293

(JMIRx Med 2025;6:e84848) doi:10.2196/84848

KEYWORDS

COVID-19 pandemic; vaccination coverage; Ecuador; immunization; routine vaccination; health disparities; vaccine hesitancy

This is the peer-review report for “Impact of the COVID-19 Pandemic on Routine Childhood Vaccination Coverage in Ecuador From 2019 to 2021: Comparative Analysis.”

Round 1 Review

General Comments

This manuscript [1] presents a thorough and compelling analysis of the impact of the COVID-19 pandemic on routine and COVID-19 vaccination coverage in Ecuador. Using national and regional data from 2019 to 2021, the authors provide clear evidence of declining immunization rates across a range of vaccines, emphasizing the public health implications of disrupted vaccination services. The use of comparative and trend analyses, as well as spatial disaggregation by region, strengthens the study’s findings. The paper is timely, well-organized, and contributes valuable insights for health system resilience in the face of future public health crises.

Specific Comments

Major Comments

1. Clarity on methodology: The study uses observational comparative analysis and descriptive statistics but would benefit from additional details on the specific statistical tests used (eg, Joinpoint regression) and any confidence intervals or measures of significance included.

2. Policy and programmatic implications: While the discussion clearly outlines the negative impact on vaccination coverage, the manuscript could be strengthened by offering more specific recommendations for public health policy, especially regarding catch-up campaigns or digital infrastructure improvements to track immunization.

3. Sociodemographic context: The analysis highlights disparities but could be improved by integrating more granular sociodemographic information (eg, income, ethnicity, rurality) to provide a deeper understanding of inequities in coverage and guide targeted interventions.

Minor Comments

4. Language and style: The manuscript would benefit from light editing to improve flow and reduce minor typographical and grammatical errors.

5. Figure/table integration: Tables are rich in data, but would be more useful if the text referenced key figures and included short interpretation notes to help readers navigate large data points.

6. Redundancy in the Introduction: Some repetition in the early paragraphs could be streamlined to maintain reader engagement.

Conflicts of Interest

None declared.

Reference

1. Sanchez J, Rodriguez AA, Cuello KPM. Impact of the COVID-19 Pandemic on Routine Childhood Vaccination Coverage in Ecuador From 2019 to 2021: Comparative Analysis. JMIRx Med 2025;6:e75293. [doi: 10.2196/75293]

Edited by A Grover; submitted 25.09.25; this is a non–peer-reviewed article;accepted 24.09.25; published 17.10.25.

Please cite as:
Adekola A
Peer Review of "Impact of the COVID-19 Pandemic on Routine Childhood Vaccination Coverage in Ecuador From 2019 to 2021: Comparative Analysis"
JMIRx Med 2025;6:e84848
URL: https://xmed.jmir.org/2025/1/e84848
doi:10.2196/84848

Peer Review of “Impact of the COVID-19 Pandemic on Routine Childhood Vaccination Coverage in Ecuador From 2019 to 2021: Comparative Analysis”

Ziqing Wang

Columbia University, New York, NY, United States

Related Articles:

Companion article: https://www.medrxiv.org/content/10.1101/2025.03.26.25324742v1

Companion article: https://med.jmirx.org/2025/1/e84851

Companion article: https://med.jmirx.org/2025/1/e75293

(JMIRx Med 2025;6:e84847) doi:10.2196/84847

KEYWORDS

COVID-19 pandemic; vaccination coverage; Ecuador; immunization; routine vaccination; health disparities; vaccine hesitancy

This is the peer-review report for “Impact of the COVID-19 Pandemic on Routine Childhood Vaccination Coverage in Ecuador From 2019 to 2021: Comparative Analysis.”

Round 1 Review

General Comments

This paper [1] presents some descriptive statistics on routine childhood vaccination coverage before and after the COVID-19 pandemic started in Ecuador. The authors found an overall declining trend in routine childhood vaccination coverage since the pandemic started. The aim of the paper is important, but the manuscript seems incomplete. For example, the authors mentioned geographical disparities in access to vaccines but did not clearly present relevant data analysis results to support the statements. Moreover, the authors mentioned that COVID-19 vaccination coverage in the general population was assessed, but they did not include such information in the manuscript.

Specific Comments

Major Comments

1. Please complete the manuscript by adding the results and interpretation of the Joinpoint regression analyses. The authors claimed that Joinpoint regression analyses were conducted but did not present and discuss the results. More importantly, please note that the Joinpoint analysis requires at least 7 time points. The authors only included 3 time points (2019, 2020, 2021). I suggest either calculating vaccination coverage percentages for at least 7 years to run the Joinpoint analysis or just presenting the descriptive statistics for each year without using the Joinpoint analysis.

2. There is a figure that plots the vaccination coverage rates in 2019, 2020, and 2021, but the authors did not provide any description or interpretation of the figure.

3. Please add descriptions for all tables and figures.

4. Please be more specific in the Data Analysis section; for example, please clearly mention what was meant by trend analysis and comparative analysis and include any specific descriptive summaries and/or statistical tests you used.

5. Please improve the organization of the Results section. For example, the regional disparities were discussed at the end of the section, yet they were presented in the first table.

6. Please narrow the focus of the manuscript. It seems that the authors aim to characterize the changes in routine childhood vaccination before and after the COVID pandemic, but in the manuscript, the authors also mention the disparities in getting the COVID-19 vaccine among the entire Ecuador population. These seem like relatively separate topics and could possibly be studied in two manuscripts.

7. Please support all claims with data or citations. For example, if the authors decide to also study the disparities in COVID-19 vaccine access, please include relevant data analysis results in the manuscript.

Minor Comments

8. At the start of the Data Analysis section, please cite the specific software used.

9. I was wondering if there is data from after the pandemic (2022 and beyond), so the authors can examine whether routine childhood vaccination coverage went back up or kept declining.

10. Please clarify what Table 1 presents and why it is included.

Conflicts of Interest

None declared.

Reference

1. Sanchez J, Rodriguez AA, Cuello KPM. Impact of the COVID-19 Pandemic on Routine Childhood Vaccination Coverage in Ecuador From 2019 to 2021: Comparative Analysis. JMIRx Med 2025;6:e75293. [doi: 10.2196/75293]

Edited by A Grover; submitted 25.09.25; this is a non–peer-reviewed article;accepted 25.09.25; published 17.10.25.

Please cite as:
Wang Z
Peer Review of "Impact of the COVID-19 Pandemic on Routine Childhood Vaccination Coverage in Ecuador From 2019 to 2021: Comparative Analysis"
JMIRx Med 2025;6:e84847
URL: https://xmed.jmir.org/2025/1/e84847
doi:10.2196/84847

Peer Review of “Impact of the COVID-19 Pandemic on Routine Childhood Vaccination Coverage in Ecuador From 2019 to 2021: Comparative Analysis”

Busurat Adenike Mudashiru

Ohio University, Athens, OH, United States

Related Articles:

Companion article: https://www.medrxiv.org/content/10.1101/2025.03.26.25324742v1

Companion article: https://med.jmirx.org/2025/1/e84851

Companion article: https://med.jmirx.org/2025/1/e75293

(JMIRx Med 2025;6:e84849) doi:10.2196/84849

KEYWORDS

COVID-19 pandemic; vaccination coverage; Ecuador; immunization; routine vaccination; health disparities; vaccine hesitancy

This is the peer-review report for “Impact of the COVID-19 Pandemic on Routine Childhood Vaccination Coverage in Ecuador From 2019 to 2021: Comparative Analysis.”

Round 1 Review

General Comments

This paper [1] is highly impactful. The authors address a significant gap in current knowledge of the impact of COVID-19 on vaccination coverage.

Specific Comments

There is a formatting issue; the figures and tables should be neatly placed.

Major Comments

1. The tables and figures should be well-labeled and referenced.

2. The limitations of the study are briefly mentioned.

3. The statistical methods should be well-presented.

Minor Comments

4. The Methods should be more explanatory.

Conflicts of Interest

None declared.

Reference

1. Sanchez J, Rodriguez AA, Cuello KPM. Impact of the COVID-19 Pandemic on Routine Childhood Vaccination Coverage in Ecuador From 2019 to 2021: Comparative Analysis. JMIRx Med 2025;6:e75293. [doi: 10.2196/75293]

Edited by A Grover; submitted 25.09.25; this is a non–peer-reviewed article;accepted 25.09.25; published 17.10.25.

Please cite as:
Mudashiru BA
Peer Review of “Impact of the COVID-19 Pandemic on Routine Childhood Vaccination Coverage in Ecuador From 2019 to 2021: Comparative Analysis”
JMIRx Med 2025;6:e84849
URL: https://xmed.jmir.org/2025/1/e84849
doi:10.2196/84849

Peer Review of "Assessing the Limitations of Large Language Models in Clinical Practice Guideline–Concordant Treatment Decision-Making on Real-World Data: Retrospective Study"

Reenu Singh

Indian Institute of Management Mumbai, Vihar Lake Rd, Mumbai, India

Related Articles:

Companion article: http://preprints.jmir.org/preprint/74899

Companion article: https://med.jmirx.org/2025/1/e84173

Companion article: https://med.jmirx.org/2025/1/e74899

(JMIRx Med 2025;6:e84175) doi:10.2196/84175

KEYWORDS

large language model; foundation model; reasoning model; treatment decision-making; aortic stenosis; clinical practice guidelines; medical data processing

This is the peer-review report for "Assessing the Limitations of Large Language Models in Clinical Practice Guideline–Concordant Treatment Decision-Making on Real-World Data: Retrospective Study."

Round 1 Review

Specific Comments

Major Comments

1. To improve the discussion on bias in large language models (LLMs) for clinical decision-making, the study [1] should include the following aspects:

If LLMs are trained predominantly on Western medical literature or specific demographic groups, their recommendations may not generalize well to diverse patient populations. If the data used to fine-tune the model lack representation from certain ethnic, gender, or socioeconomic groups, the artificial intelligence may produce recommendations that are not universally applicable. Even with a diverse dataset, biases can arise due to model architecture, reinforcement learning strategies, or human-in-the-loop feedback mechanisms that shape model responses.

2. What datasets were used? If real patient data were used, specify its source (eg, electronic health records, clinical trial data, or synthetic datasets). Provide the total number of cases or records used for testing the LLMs. If synthetic data were generated, describe the method used to create the data. Were diverse age groups, genders, and ethnic backgrounds represented? A lack of diversity in data can affect the generalizability of results.

3. What datasets were used? If real patient data were used, specify its source (eg, electronic health records, clinical trial data, or synthetic datasets). Provide the total number of cases or records used for testing the LLMs. If synthetic data were generated, describe the method used to create the data. Were diverse age groups, genders, and ethnic backgrounds represented? A lack of diversity in data can affect the generalizability of results.

The study's impact can be significantly enhanced by addressing the following challenges: Raw medical reports often include free-text narratives, physician notes, abbreviations, and inconsistencies, requiring advanced natural language processing techniques such as entity recognition, text normalization, and standardization. These reports may also contain irrelevant information, redundancies, or nonessential clinical details. Effective preprocessing is essential to filter out unnecessary content while preserving critical medical insights. A key consideration is how to optimize this preprocessing to mitigate these challenges efficiently.

4. The study's impact can be significantly enhanced by addressing the following challenges: Raw medical reports often include free-text narratives, physician notes, abbreviations, and inconsistencies, requiring advanced natural language processing techniques such as entity recognition, text normalization, and standardization. These reports may also contain irrelevant information, redundancies, or nonessential clinical details. Effective preprocessing is essential to filter out unnecessary content while preserving critical medical insights. A key consideration is how to optimize this preprocessing to mitigate these challenges efficiently.

Round 2 Review

1. The authors have addressed the comments satisfactorily.

Conflicts of Interest

None declared.

Reference

1. Roeschl T, Hoffmann M, Hashemi D, et al. Assessing the limitations of large language models in clinical practice guideline–concordant treatment decision-making on real-world data: retrospective study. JMIRx Med 2025;6:e84173. [doi: 10.2196/84173]

Abbreviations

LLM: large language model

Edited by A Grover; submitted 15.09.25; this is a non–peer-reviewed article;accepted 15.09.25; published 03.11.25.

Please cite as:
Singh R
Peer Review of "Assessing the Limitations of Large Language Models in Clinical Practice Guideline–Concordant Treatment Decision-Making on Real-World Data: Retrospective Study"
JMIRx Med 2025;6:e84175
URL: https://xmed.jmir.org/2025/1/e84175
doi:10.2196/84175

Peer Review of "Assessing the Limitations of Large Language Models in Clinical Practice Guideline–Concordant Treatment Decision-Making on Real-World Data: Retrospective Study"

Andrej Novak

University of Zagreb, Ul. Radoslava Cimermana 88, Zagreb, Croatia

Related Articles:

Companion article: http://preprints.jmir.org/preprint/74899

Companion article: https://med.jmirx.org/2025/1/e84173

Companion article: https://med.jmirx.org/2025/1/e74899

(JMIRx Med 2025;6:e84174) doi:10.2196/84174

KEYWORDS

large language model; foundation model; reasoning model; treatment decision-making; aortic stenosis; clinical practice guidelines; medical data processing

This is the peer-review report for "Assessing the Limitations of Large Language Models in Clinical Practice Guideline–Concordant Treatment Decision-Making on Real-World Data: Retrospective Study."

Round 1 Review

The authors of this paper [1] set out to determine whether modern large language models (LLMs) can make treatment decisions for severe aortic stenosis based on uncurated, free-text clinical data in routine practice. This question addresses a significant gap in the literature: while earlier work demonstrated that LLMs could agree with expert tumor boards or heart teams when provided with highly structured or preprocessed information, the realities of clinical documentation—discharge summaries, imaging reports, and free-text notes—remain unstructured and noisy. It seems that even top LLMs fail to deliver reliable, unbiased treatment recommendations from raw clinical text. They perform well only with expert-crafted summaries and embedded guidelines, highlighting that data representation (and prompt engineering) is key.

Methods: The authors should be commended for the exceptionally detailed Methods section, which carefully notes subtleties such as each model's maximum context. Their inclusion of a simple non-LLM reference model alongside a broad spectrum of open- and closed-source LLMs represents thoughtful experimental design, and their 4-arm RAW→RAW+→SUM→SUM+ framework neatly isolates the impact of data representation and guideline context.

Analysis and Results: Using Cohen κ to assess agreement on a binary decision task is appropriate; when paired with accuracy, it gives a fuller picture than raw percentages alone. Supplementing κ with intraclass correlation coefficients (ICCs) and normalized Shannon entropy to gauge reliability across repeated runs is also sound. The results themselves are compelling. Across 9 models, κ values on raw text ranged from slight negative agreement up to only fair (–0.47 to 0.22), and ICCs were poor, demonstrating that without curated input, even leading LLMs can not reliably distinguish surgical aortic valve replacement from transcatheter aortic valve replacement. When expert summaries plus guideline text were provided, κ jumped into the moderate - substantial range (up to 0.63) and ICCs reached good–excellent levels. That consistent, monotonic improvement from RAW→RAW+→SUM→SUM+ (replicated across open and closed models) makes a strong, convincing case that data representation, not just model capability, drives performance.

That said, the retrospective, single-center design with only 80 cases further constrains generalizability; patient populations and documentation styles vary widely across institutions. The way indeterminate recommendations were handled in metrics (counted as "wrong" for κ and accuracy, but excluded from bias calculations) may also skew the interpretation of model caution versus error. Finally (as noted in the Limitations), on a philosophical level—treating heart-team decisions, which are themselves subjective, as infallible ground truth risks overstating LLM shortcomings.

Beyond the major strengths and limitations previously discussed, I have identified several minor points that would further strengthen the manuscript:

1. The format and provenance of the SUM ("case summary") reports require clearer specification. Although the authors note these summaries were "manually generated," it would be helpful to state whether they followed a standardized template, who exactly drafted them (eg, experienced cardiologists, research

assistants), and which elements of the Heart Team protocol they distilled into each summary.

2. The authors report that the original medical documents were saved as PDFs and later converted to plain text. It would be helpful to clarify this process to avoid confusion, since LLMs accessed via chat interfaces or application programming interfaces often struggle with PDF inputs or text embedded in images, treating them differently from pure text. A brief discussion acknowledging this limitation—and explaining how PDF parsing was handled or validated—would help readers assess real-world applicability.

3. Raw inputs (PDFs and summaries) were provided in German (except for BioGPT, which required translation to English). A comment in the Discussion about how model performance can vary by input language—perhaps citing studies that showed different results in Polish versus English—would contextualize the findings for non-English clinical settings:

- Rosoł M, Gąsior JS, Łaba J, Korzeniewski K, Młyńczak M. Evaluation of the performance of GPT-3.5 and GPT-4 on the Polish Medical Final Examination. *Sci Rep*. 2023;13(1):20512.

4. The Discussion section feels comparatively weak and could be strengthened by broader literature coverage. For instance, a brief discussion of input formats—pure text versus multimodal inputs—would be valuable, especially given the inclusion of GPT-4o, which handles images. Preliminary studies in this area include:

- Günay et al. Comparison of emergency medicine specialist, cardiologist, and ChatGPT in electrocardiography assessment. *Am J Emerg Med*. 2024 Jun;80:51-60.
- Zeljkovic et al. Beyond text: the impact of clinical context on GPT-4's 12-lead electrocardiogram interpretation accuracy. *Canadian J Cardiol*. 2025 Jul;41(7):1406-1414.

These compare electrocardiogram interpretation with and without accompanying clinical context and demonstrate the importance of textual input alongside images.

It would also be helpful to reference work showing that, despite similar hallucination tendencies, LLMs perform strongly on standardized exams, for example:

- Gilson et al. How does ChatGPT perform on the USMLE? Implications for medical education and knowledge assessment. *JMIR Med Educ*. 2023 Feb 8;9:e45312.
- Novak et al. The pulse of artificial intelligence in cardiology: evaluating state-of-the-art LLMs for clinical cardiology. *medRxiv*. Preprint posted online on January 30, 2024.

These additions could situate the findings within a broader context of multimodal and high-stakes assessment.

5. As an exploratory aside, it would be interesting to evaluate how the newest reasoning-focused models (eg, "o3" or "o4") perform on this task. Although this is likely beyond the current scope, including a sentence to that effect in the manuscript's Limitations section could guide future research.

6. For consistency and precision, when describing model access in the "Large Language Models" section (and elsewhere in the text), the manuscript should explicitly cite the exact supplementary tables or materials (eg, "see Table S1 for model details and context sizes") rather than referring generically to "the Supplementary."

7. In the Statistical Methods subsection, rather than stating that nonnormally distributed data were compared using the Mann-Whitney *U* test "for nonnormally distributed continuous variables," the phrasing could be tightened to "for variables departing from normality" or "for variables not following a normal distribution" to align with standard statistical terminology.

Conflicts of Interest

None declared.

Reference

1. Roeschl T, Hoffmann M, Hashemi D, et al. Assessing the limitations of large language models in clinical practice guideline–concordant treatment decision-making on real-world data: retrospective study. JMIRx Med 2025;6:e84173. [doi: 10.2196/84173]

Abbreviations

ICC: intraclass correlation coefficient
LLM: large language model

Edited by A Grover; submitted 15.09.25; this is a non–peer-reviewed article;accepted 15.09.25; published 03.11.25.

Please cite as:
Novak A
Peer Review of "Assessing the Limitations of Large Language Models in Clinical Practice Guideline–Concordant Treatment Decision-Making on Real-World Data: Retrospective Study"
JMIRx Med 2025;6:e84174
URL: https://xmed.jmir.org/2025/1/e84174
doi:10.2196/84174

Peer Review of "Evaluating the Financial Factors Influencing Maternal, Newborn, and Child Health in Africa: Tobit Regression and Data Envelopment Analysis"

Masoud Mahundi

University of Dar es Salaam, Dar es Salaam, United Republic of Tanzania

Related Articles:

Companion article: https://www.medrxiv.org/content/10.1101/2024.02.22.24303217v1

Companion article: https://med.jmirx.org/2025/1/e85578

Companion article: https://med.jmirx.org/2025/1/e59703

(JMIRx Med 2025;6:e85383) doi:10.2196/85383

KEYWORDS

financial determinants; maternal, newborn, and child health; health care efficiency; Africa; health expenditure; data envelopment analysis; Tobit regression

This is the peer-review report for "Evaluating the Financial Factors Influencing Maternal, Newborn, and Child Health in Africa: Tobit Regression and Data Envelopment Analysis."

Round 1 Review

General Comments

Although the title might initially seem misleading, the paper [1] tackles an essential issue of efficiency in delivering maternal, newborn, and child health services in Africa. Given the well-known challenges facing maternal and newborn health in the region, the importance of this study cannot be overstated.

Specific Comments

Major Comments

1. The whole abstract is more about general efficiency in health care systems and less about the financial factors influencing maternal, newborn, and child health (MNCH) in Africa. The title has to reflect what the study actually presents.
2. If the key aim of the study was to determine how financial factors influence MNCH, one would then expect to see, in the abstract, the extent of the influence of financial factors such as health expenditure, coverage index, and expenditure per capita.
3. The Introduction section starts with a presentation of the number of women dying in 2020 and the number of children dying in 2021. These data are not supported with any citations. It is also a little strange that for the number of women dying, the study refers to 2020 data, while for that of children, the study refers to 2021.
4. The Introduction section does not provide sufficient motivation for investigating efficiency in health systems. The first paragraph presents the maternal health challenges, while the second paragraph quickly goes to the methods for establishing efficiency. There is no connection as to why investigating efficiency is necessary.
5. The Introduction section provides a descriptive review of other studies without depth. It lists the different studies without synthesizing them. It would be useful if they were at least lifted up to present issues/themes so it is easy to connect with what the study is about.
6. There is a sentence in the Methods section (Data Sources and Variables) that says "Input, output, and explanatory variables were selected to assess the accuracy of the WHO [World Health Organization]..." Are there three types of variables in your study?
7. What is presented as stages of data envelopment analysis does not go further to describe how the study made use of these stages. Much of the presentations are about what these stages are and sometimes the historical background. It would be useful to put more emphasis on how the study used these stages so it assures the credibility and reliability of the findings.
8. The presented results do not have a clear foundation from the methods. The chain of evidence from the data is lacking, from their processing to their results.
9. The parameters presenting the results are not clearly defined. It says "...26% with a score of 1." There is no proper introduction of the ranges for a reader to comprehend the meaning of a score of 1. It also states that "...average efficiency score (TE-VRS) across all countries is 0.849 for VRS [variable returns to scale]."
10. The Discussion section needs revising. It does not directly connect to the findings of the study, despite the challenges of the results. The Discussion section further presents a couple of statistics, especially in the first and second

paragraphs, without sources or a clear connection to the findings.

Conflicts of Interest

None declared.

Reference

1. Er-Rays Y, M'dioud M, Ait-Lemqeddem H, El Moutaqi B. Evaluating the financial factors influencing maternal, newborn, and child health in Africa: Tobit regression and data envelopment analysis. JMIRx Med 2025;6:e59703. [doi: 10.2196/59703]

Edited by F Wu; submitted 06.10.25; this is a non–peer-reviewed article;accepted 06.10.25; published 28.11.25.

Please cite as:
Mahundi M
Peer Review of "Evaluating the Financial Factors Influencing Maternal, Newborn, and Child Health in Africa: Tobit Regression and Data Envelopment Analysis"
JMIRx Med 2025;6:e85383
URL: https://xmed.jmir.org/2025/1/e85383
doi:10.2196/85383

Peer Review of "Evaluating the Financial Factors Influencing Maternal, Newborn, and Child Health in Africa: Tobit Regression and Data Envelopment Analysis"

Titilayo Deborah Olorunyomi

Related Articles:

Companion article: https://www.medrxiv.org/content/10.1101/2024.02.22.24303217v1

Companion article: https://med.jmirx.org/2025/1/e85578

Companion article: https://med.jmirx.org/2025/1/e59703

(JMIRx Med 2025;6:e85382) doi:10.2196/85382

KEYWORDS

financial determinants; maternal, newborn, and child health; health care efficiency; Africa; health expenditure; data envelopment analysis; Tobit regression

This is the peer-review report for "Evaluating the Financial Factors Influencing Maternal, Newborn, and Child Health in Africa: Tobit Regression and Data Envelopment Analysis."

Round 1 Review

The manuscript [1] provides a valuable contribution to the understanding of health care system efficiency in Africa, particularly in the context of maternal, newborn, and child health. The study is ethical, with appropriate use of data from reputable sources such as the World Health Organization, and the methods employed—data envelopment analysis and Tobit regression—are suitable for assessing the technical efficiency of health care systems across 46 African countries.

The material is original, and the paper addresses a significant gap in the literature by focusing on the financial and efficiency factors impacting maternal, newborn, and child health in Africa. Related work is discussed and cited adequately, although a few more recent studies could be included to strengthen the literature review.

The writing is generally clear, though there are some areas where the discussion of the results could benefit from more detail. The study methods are appropriate for the research objectives, and the data used appear to be valid and reliable. The findings are significant and present actionable insights for policymakers, especially in terms of understanding the inefficiencies in the health care systems that impact maternal, newborn, and child health outcomes.

The conclusions are reasonable and are supported by the data, although more detailed recommendations for practical application could enhance the paper's impact. The topic is certainly of interest to the readership, as it addresses key issues surrounding health care efficiency and the achievement of Sustainable Development Goal 3 in Africa.

Overall, I recommend the manuscript for publication with minor revisions to improve the clarity of some sections and provide more detailed policy recommendations.

Conflicts of Interest

None declared.

Reference

1. Er-Rays Y, M'dioud M, Ait-Lemqeddem H, El Moutaqi B. Evaluating the financial factors influencing maternal, newborn, and child health in Africa: Tobit regression and data envelopment analysis. JMIRx Med 2025;6:e59703. [doi: 10.2196/59703]

Edited by F Wu; submitted 06.10.25; this is a non–peer-reviewed article;accepted 06.10.25; published 28.11.25.

Please cite as:
Olorunyomi TD
Peer Review of "Evaluating the Financial Factors Influencing Maternal, Newborn, and Child Health in Africa: Tobit Regression and Data Envelopment Analysis"
JMIRx Med 2025;6:e85382
URL: https://xmed.jmir.org/2025/1/e85382
doi:10.2196/85382

Authors' Response to Peer Reviews of "Applications of Indocyanine Green in Breast Cancer for Sentinel Lymph Node Mapping: Protocol for a Scoping Review"

Feryal Kurdi*, MD; Yahya Kurdi*, MD; Igor Vladimirovich Reshetov, MD, PhD

Department of Oncology, Radiotherapy and Plastic and Reconstructive Surgery, Sechenov University, Bolshaya Pirogovskaya, 6c1, Moscow, Russian Federation

*these authors contributed equally

Corresponding Author:
Feryal Kurdi, MD
Department of Oncology, Radiotherapy and Plastic and Reconstructive Surgery, Sechenov University, Bolshaya Pirogovskaya, 6c1, Moscow, Russian Federation

Related Articles:

Companion article: https://www.medrxiv.org/content/10.1101/2024.07.30.24311256v1

Companion article: https://med.jmirx.org/2025/1/e69705

Companion article: https://med.jmirx.org/2025/1/e66213

(JMIRx Med 2025;6:e68769) doi:10.2196/68769

KEYWORDS

indocyanine green; ICG; sentinel lymph node; breast cancer; breast; fluorescence; axillary lymph node mapping; NIR; surgical planning; near-infrared

This is the authors' response to peer-review reports for "Applications of Indocyanine Green in Breast Cancer for Sentinel Lymph Node Mapping: Protocol for a Scoping Review."

Round 1 Review

Anonymous [1]

General Comments

This paper [2] summarized the application value and existing problems of indocyanine green (ICG) in sentinel lymph node (SLN) biopsy of early breast cancer, which has positive significance for improving the accuracy of clinical SLN detection. This study has certain clinical value.

Response: Thank you for your thoughtful comments and feedback on our paper. Below are my responses to your points.

Specific Comments

Major Comments

1. Due to the high hardware requirements for the clinical application of ICG, the number of relevant studies in the search is relatively small. It is hoped that the author can search the recent, relevant literature to improve the credibility of this review.

Response: This paper is a protocol for a scoping review, serving as a roadmap for the search strategy and inclusion criteria that we will follow. As such, it outlines our plan rather than reporting the outcomes of the literature search. As noted in Multimedia Appendix 1, we will conduct a comprehensive search across multiple databases to ensure the inclusion of all relevant, recent studies.

2. It is hoped that the author will analyze and compare the advantages and disadvantages of ICG and traditional SLN biopsy methods, so as to guide clinicians to adopt appropriate methods for appropriate patients.

Response: As indicated in Multimedia Appendix 1, this comparison is a core objective of our review. We hope these clarifications address your concerns.

References

1. Anonymous. Peer review of "Applications of Indocyanine Green in Breast Cancer for Sentinel Lymph Node Mapping: Protocol for a Scoping Review. JMIRx Med 2024;5:e69705. [doi: 10.2196/69705]

2. Kurdi F, Kurdi Y, Reshetov IV. Applications of indocyanine green in breast cancer for sentinel lymph node mapping: protocol for a scoping review. JMIRx Med 2024;5:e66213. [doi: 10.2196/66213]

Abbreviations

ICG: indocyanine green
SLN: sentinel lymph node

Edited by S Tungjitviboonkun; submitted 13.11.24; this is a non–peer-reviewed article;accepted 13.11.24; published 06.01.25.

Please cite as:
Kurdi F, Kurdi Y, Reshetov IV
Authors' Response to Peer Reviews of "Applications of Indocyanine Green in Breast Cancer for Sentinel Lymph Node Mapping: Protocol for a Scoping Review"
JMIRx Med 2025;6:e68769
URL: https://xmed.jmir.org/2025/1/e68769
doi:10.2196/68769

Authors' Response to Peer Reviews of "The Impact of Rural Alimentation on the Motivation and Retention of Indigenous Community Health Workers in India: Qualitative Study"

Ajit Kerketta*, MHA; Raghavendra A N*, PhD

CHRIST (Deemed to be University), Hosur Road, Bhavani Nagar, Bengaluru, India

*all authors contributed equally

Corresponding Author:
Ajit Kerketta, MHA
CHRIST (Deemed to be University), Hosur Road, Bhavani Nagar, Bengaluru, India

Related Articles:

Companion article: https://www.medrxiv.org/content/10.1101/2023.04.12.23288461v1

Companion article: https://med.jmirx.org/2025/1/e70808

Companion article: https://med.jmirx.org/2025/1/e48346

(JMIRx Med 2025;6:e70059) doi:10.2196/70059

KEYWORDS

rural alimentation; community health workers; motivation; retention; rural health; rural nutrition; workforce

This is the authors' response to peer-review reports for "The Impact of Rural Alimentation on the Motivation and Retention of Indigenous Community Health Workers in India: Qualitative Study."

Round 1 Review [1]

General Comments

This paper [2] has given the impression that the researcher has done thorough homework before starting the research and it is evident in the paper. Case methodology and thematic analysis are a few of the approaches that depict the quality of the paper. Overall, as a reviewer, it is my opinion that the research paper is of quality.

Specific Comments

1. A few more factors like government initiatives should be included in studying the impact on the motivation and retention of community health workers.

Response: Factors such as government initiatives and policies have been additionally incorporated into the Discussion section.

Major Comments

1. I feel that the analysis also can include education as a parameter.

2. The thematic analysis is one of the strengths of this research and is appreciated.

Response: Due to time constraints, education could not be included as a sample parameter.

Minor Comments

1. Common wording should be used in every section of the paper, like qualitative case research methodology and qualitative case research.

Response: The term "qualitative case research" has been consistently used throughout the study.

References

1. Kumar Thalari S. Peer review of "The Impact of Rural Alimentation on the Motivation and Retention of Indigenous Community Health Workers in India: Qualitative Study". JMIRx Med 2025;6:e70808. [doi: 10.2196/70808]
2. Kerketta A, A N R. The impact of rural alimentation on the motivation and retention of Indigenous community health workers in India: qualitative study. JMIRx Med 2025;6:e48346. [doi: 10.2196/48346]

Edited by A Schwartz; submitted 13.12.24; this is a non–peer-reviewed article;accepted 13.12.24; published 23.01.25.

Please cite as:
Kerketta A, A N R
Authors' Response to Peer Reviews of "The Impact of Rural Alimentation on the Motivation and Retention of Indigenous Community Health Workers in India: Qualitative Study"
JMIRx Med 2025;6:e70059
URL: https://xmed.jmir.org/2025/1/e70059
doi:10.2196/70059

Author's Response to Peer Reviews of "Identifying Safeguards Disabled by Epstein-Barr Virus Infections in Genomes From Patients With Breast Cancer: Chromosomal Bioinformatics Analysis"

Bernard Friedenson, PhD

Department of Biochemistry and Medical Genetics, Cancer Center, University of Illinois Chicago, 900 s Ashland, Chicago, IL, United States

Corresponding Author:
Bernard Friedenson, PhD
Department of Biochemistry and Medical Genetics, Cancer Center, University of Illinois Chicago, 900 s Ashland, Chicago, IL, United States

Related Articles:

Companion article: https://www.medrxiv.org/content/10.1101/2023.07.03.23292185v1

Companion article: https://med.jmirx.org/2025/1/e70039

Companion article: https://med.jmirx.org/2025/1/e70041

Companion article: https://med.jmirx.org/2025/1/e50712

(JMIRx Med 2025;6:e69307) doi:10.2196/69307

KEYWORDS

breast; cancer; oncology; ovarian; virus; viral; Epstein-Barr; herpes; bioinformatics; chromosome; gene; genetic; chromosomal; DNA; genomic; BRCA; metastasis; biology

This is the author's response to peer-review reports for "Identifying Safeguards Disabled by Epstein-Barr Virus Infections in Genomes From Patients With Breast Cancer: Chromosomal Bioinformatics Analysis."

Round 1 Review

Anonymous [1]

Review Report With Major Revisions for the Paper

Title: "Herpesvirus infections eliminate safeguards against breast cancer and its metastasis: comparable to hereditary breast cancers"

Summary

The paper [2] hypothesizes that Epstein-Barr virus (EBV) infections promote breast cancer by disabling cancer safeguards. It is a bioinformatics analysis of public information from about 2100 breast cancers. The study finds that breast and ovarian cancer breakpoints cluster around EBV-associated cancer breakpoints, suggesting a significant role of EBV in promoting these cancers. The paper also identifies similarities in the molecular and cellular disruptions caused by EBV with those found in hereditary breast cancers.

Major Revisions Needed

Clarification of Hypotheses and Objectives

The hypothesis, while intriguing, needs clearer articulation. Specifically, the connection between EBV and breast cancer needs more explicit theoretical underpinning. Clarify the objectives and expected outcomes of the study at the outset.

Response: The objectives and expected outcomes of the study were clarified at the outset in the Abstract and Introduction.

Methodological Rigor and Data Sources

While the bioinformatics approach is robust, it would benefit from a more detailed description of the methods and algorithms used. Additionally, the selection criteria for the breast cancer data should be justified more thoroughly to avoid selection bias.

Response: A more detailed description of the methods and algorithms used has been added in the Methods section (page 6).

Statistical Analysis

The statistical methods used need more comprehensive detailing. For complex analyses, ensure the statistical assumptions and any transformations of data are clearly explained. Include more information on the statistical tests used for hypothesis testing and the justification for their use.

Response: I included more information on the statistical tests, the justification, and limitations of their use (page 7).

Comparative Analysis

The comparison between hereditary breast cancers and those potentially caused by EBV is insightful. However, a more detailed comparative analysis would strengthen the argument. This could include molecular or genetic profiling comparisons.

Response: I added a more detailed comparative analysis with results in Figure 2H and Table S2, as described on page 10.

Discussion on Contradictory or Supporting Evidence

The discussion section should address not only the supporting evidence but also any contradictory findings in the literature. This balance is crucial for a nuanced understanding of the subject.

Response: The paper's hypothesis more clearly accounts for the absence of demonstrable EBV infection in breast cancer, explaining contradictory results. The other contradictory result posits an imperfect palindrome on chromosome 11. This result is tested on page 13.

Implications and Future Research Directions

The implications of these findings are profound but need clearer articulation. Discuss the potential impact on breast cancer treatment and prevention strategies. Also, outline future research directions, particularly in clinical or experimental studies to confirm these bioinformatics findings.

Response: I articulated the implications of these finding more clearly with their impact on breast cancer treatment and prevention strategies. I also outlined future research directions with clinical or experimental studies to confirm the bioinformatics findings (Discussion, page 16).

References

Please add more background information about breast cancer (please cite: 1. Cao Y, Efetov S, He M, et al. Updated clinical perspectives and challenges of chimeric antigen receptor-T cell therapy in colorectal cancer and invasive breast cancer. Arch Immunol Ther Exp (Warsz). Aug 11, 2023;71(1):19. [doi: 10.1007/s00005-023-00684-x] [Medline: 37566162]; and 2. Liu Y, Lu S, Sun Y, et al. Deciphering the role of QPCTL in glioma progression and cancer immunotherapy. Front Immunol. Mar 29, 2023;14:1166377. [doi: 10.3389/fimmu.2023.1166377] [Medline: 37063864]).

Response: I added these references.

Concluding Remarks

The paper presents a novel and potentially significant hypothesis linking EBV to breast cancer. However, it requires major revisions to enhance its methodological rigor, clarity, and comprehensiveness. Addressing these concerns will significantly strengthen the manuscript's impact and contribution to the field.

Anonymous [3]

Dear Author,

After a thorough review of the paper titled "Herpesvirus infections eliminate safeguards against breast cancer and its metastasis: comparable to hereditary breast cancers" by Bernard Friedenson, here is the negative feedback and evaluation, along with a recommendation for the inclusion of a specific article in the discussion section.

Negative Feedback and Evaluation

Clarity and Scope

The paper ambitiously attempts to link Epstein-Barr virus (EBV) infections to breast cancer development and metastasis. While the hypothesis is intriguing, the narrative sometimes lacks clarity and could benefit from a more focused scope. The vast amount of data and the complex mechanisms presented can be overwhelming and occasionally detract from the main message.

Response: I focused the scope in this revision in the Abstract and Introduction.

Methodological Concerns

The reliance on bioinformatics analyses and previously published datasets raises questions about the direct experimental validation of the proposed mechanisms. Although the computational approach is valid, the absence of direct experimental evidence or validation in breast cancer samples limits the strength of the conclusions.

Response: I explained in the Discussion section that direct experimental evidence or validation has already been done. EBV-infected human mammary epithelial cells produce breast cancer in immunosuppressed mice (page 17).

Interpretation of Data

The interpretation of viral homology and its impact on cancer development is speculative in several sections. The connections made between EBV infections, chromosomal breakpoints, and cancerous mutations rely heavily on correlative data without sufficient causal evidence. A more cautious interpretation of the results, highlighting the need for further experimental validation, would strengthen the manuscript.

Response: I added more evidence (Figure 2H and Table S2) to the association of EBV infection and cancer development and took greater care throughout to interpret the results more cautiously.

Consideration of Alternate Hypotheses

The paper could benefit from a more balanced discussion of alternative hypotheses explaining the observed data. For instance, the role of other environmental, genetic, or lifestyle factors in breast cancer development is not adequately considered. Acknowledging and discussing these potential confounders would provide a more comprehensive understanding of the complex etiology of breast cancer.

Response: I explained how EBV relates to alternate hypotheses and exacerbates the effects of other known breast cancer risk factors (page 16).

References and Current Literature

While the paper cites a significant amount of relevant literature, it sometimes overlooks recent studies that could either support or challenge the proposed hypotheses. Incorporating a more

current and diverse range of references would enhance the paper's relevance and credibility.

Response: I included more information from more current and diverse ranges of references.

Recommendation for Discussion Inclusion

To broaden the discussion and contextualize the findings within the broader research landscape, it is recommended to include the following article in the discussion section.

Al-Awaida W, Al-Ameer HJ, Sharab A, Akasheh RT. Modulation of wheatgrass (Triticum aestivum Linn) toxicity against breast cancer cell lines by simulated microgravity. Curr Res Toxicol. Sep 19, 2023;5:100127. [doi: 10.1016/j.crtox.2023.100127] [Medline: 37767028]

Incorporating this article could provide valuable insights into innovative approaches for studying cancer therapies. Specifically, the effects of simulated microgravity on the efficacy of natural compounds like wheatgrass against breast cancer could open up new avenues for research on the environmental and physical conditions affecting cancer treatment outcomes. Discussing this study would enrich the manuscript by introducing the concept of microgravity as a novel factor influencing cancer cell behavior and therapy resistance, thereby offering a broader perspective on cancer research methodologies and therapeutic strategies.

Response: I could not find a way to apply and cite this interesting work since it was so far afield from the manuscript.

Round 2 Review

Anonymous [3]

General Comments

This paper tests the idea that EBV infections can help cause breast cancer by weakening the body's defenses against cancer. The study uses bioinformatics to compare chromosome breakpoints in breast cancer to those in cancers known to be caused by EBV. The results show that EBV might play a role in breast cancer by damaging important cell functions.

Specific Comments

Major Comments

The methods section needs more details about how the datasets were chosen and combined.

Response: More details on how the datasets were chosen have been added.

The discussion should explain more about how EBV might cause the chromosome breaks and rearrangements seen in breast cancer.

Response: The discussion includes an expanded explanation about how EBV might cause the chromosome breaks and rearrangements seen in breast cancer.

More data or references are needed to support the idea that EBV helps breast cancer spread to other parts of the body.

Response: A new Figure 7 and more data have been added. Additional references have also been added, and the metastasis topic has been clarified and expanded.

Minor Comments

Adding more references would strengthen the sections that talk about how EBV affects breast cancer.

Response: Many more references have been added.

Figures and tables should be clearly mentioned in the text to help readers follow the data.

Response: Figures and tables are now more prominently mentioned in the text.

Some parts of the manuscript need clearer writing and better organization, especially where complex bioinformatics results are explained.

Response: I revised the manuscript with clearer writing and better organization, especially where complex bioinformatics results are explained.

The abstract should be revised to clearly highlight the main findings and why they are important.

Response: I revised the Abstract to highlight the main findings and why they are important.

Make sure all abbreviations are defined when they are first used to help readers understand the text better.

Response: I went through the manuscript to be sure all abbreviations were defined. I also added a glossary containing abbreviations, gene names, and viruses.

References

1. Anonymous. Peer review of "Identifying Safeguards Disabled by Epstein-Barr Virus Infections in Genomes From Patients With Breast Cancer: Chromosomal Bioinformatics Analysis". JMIRx Med 2025;6:e70039. [doi: 10.2196/70039]
2. Friedenson B. Identifying safeguards disabled by Epstein-Barr virus infections in genomes from patients with breast cancer: chromosomal bioinformatics analysis. JMIRx Med 2025;6:e50712. [doi: 10.2196/50712]
3. Anonymous. Peer review of "Identifying Safeguards Disabled by Epstein-Barr Virus Infections in Genomes From Patients With Breast Cancer: Chromosomal Bioinformatics Analysis". JMIRx Med 2025;6:e70041. [doi: 10.2196/70041]

Abbreviations

EBV: Epstein-Barr virus

Edited by A Schwartz; submitted 26.11.24; this is a non–peer-reviewed article;accepted 26.11.24; published 29.01.25.

Please cite as:
Friedenson B
Author's Response to Peer Reviews of "Identifying Safeguards Disabled by Epstein-Barr Virus Infections in Genomes From Patients With Breast Cancer: Chromosomal Bioinformatics Analysis"
JMIRx Med 2025;6:e69307
URL: https://xmed.jmir.org/2025/1/e69307
doi:10.2196/69307

Authors' Response to Peer Reviews of "Mothers' Knowledge of and Practices Toward Oral Hygiene of Children Aged 5-9 Years in Bangladesh: Cross-Sectional Study"

Tahazid Tamannur[1], MPH; Sadhan Kumar Das[1], MPH; Arifatun Nesa[2], MPH; Foijun Nahar[1], MPH; Nadia Nowshin[1], MPH; Tasnim Haque Binty[1], MPH; Shafiul Azam Shakil[2], MPH; Shuvojit Kumar Kundu[3], MPH; Md Abu Bakkar Siddik[4], MPH; Shafkat Mahmud Rafsun[5], MPH; Umme Habiba[6], MPH; Zaki Farhana[7], MS; Hafiza Sultana[1], MPhil; Anton Abdulbasah Kamil[8], PhD; Mohammad Meshbahur Rahman[9], MS

[1]Department of Health Education, National Institute of Preventive and Social Medicine, Dhaka, Bangladesh
[2]Department of Public Health and Hospital Administration, National Institute of Preventive and Social Medicine, Mohakhali, Dhaka, Bangladesh
[3]Directorate General of Health Services, Ministry of Health & Family Welfare, Government of the People's Republic of Bangladesh, Dhaka, Bangladesh
[4]School of the Environment, Nanjing University, Nanjing, China
[5]Dental Speciality Center, Dhaka, Bangladesh
[6]BRAC James P Grant School of Public Health, BRAC University, Dhaka, Bangladesh
[7]Credit Information Bureau, Bangladesh Bank, Dhaka, Bangladesh
[8]Department of Business Administration, Istanbul Gelisim University, Istanbul, Turkey
[9]Department of Biostatistics, National Institute of Preventive and Social Medicine, Dhaka, Bangladesh

Corresponding Author:
Mohammad Meshbahur Rahman, MS
Department of Biostatistics, National Institute of Preventive and Social Medicine, Dhaka, Bangladesh

Related Articles:

Companion article: https://www.medrxiv.org/content/10.1101/2024.04.05.24305403v1

Companion article: https://med.jmirx.org/2025/1/e70142

Companion article: https://med.jmirx.org/2025/1/e70144

Companion article: https://med.jmirx.org/2025/1/e59379

(JMIRx Med 2025;6:e70145) doi:10.2196/70145

KEYWORDS

mothers' knowledge and practices; oral hygiene; child oral health; Bangladesh

This is the authors' response to peer-review reports for "Mothers' Knowledge of and Practices Toward Oral Hygiene of Children Aged 5-9 Years in Bangladesh: Cross-Sectional Study."

Round 1 Review

Reviewer BZ [1]

This is an interesting piece of research [2], which highlights mothers' knowledge and practices regarding their children's oral health in Dhaka City. However, several issues made the study scientifically questionable. The major issues are as follows. The study included mothers from two hospitals in Dhaka City, but the title of the study does not mention this. The sample selection from the mothers visiting the hospitals might not represent general mothers from the whole of Dhaka. Thus, this study might not be generalizable to all mothers in Dhaka City.

Response: The authors are grateful to the reviewers for critically reviewing our manuscript. We agree with the comments. Respondents of this study were the mothers visiting the tertiary-level hospitals of Dhaka City. Generally, the respondents visiting hospitals belonged to all administrative wards (small regions of Dhaka), and it is convenient to get the mothers with children aged 5 - 9 years to interview. That is why we chose tertiary-level hospitals to reach the respondents. However, we revised our manuscript title and omitted "Dhaka" from the title. The new title is "Knowledge and practices towards oral hygiene of children aged 5 - 9 years old: a cross-sectional study among mothers visited tertiary level hospitals."

Introduction

Revise the last paragraph of the Introduction to highlight the study gap in Bangladesh and clearly state the objective of the

study. Use the formal word "mother" and avoid the word "moms."

Response: We appreciate the reviewer for this comment. We revised the Introduction of our study and replaced the word "Moms" with mother.

Methods

Study Setting and Participants

Give clear reasoning as to why you selected study participants from the hospitals. The last line is confusing. It is not clear whether the participants filled out the questionnaire on their own or they were interviewed by the enumerators.

Response: We are thankful to the reviewer for this comment. Respondents of this study were the mothers visiting the tertiary-level hospitals of Dhaka City. Generally, the respondents who visited hospitals belonged to all administrative wards (small regions of Dhaka), and it was convenient to get this group of mothers with children aged 5 - 9 years to interview. That is why we chose tertiary-level hospitals to reach the respondents. However, we revised our manuscript title and omitted "Dhaka" from the title. We interviewed the respondents, and the sentence was revised in our revised manuscript.

Sampling Technique

Please mention the nonresponse bias for the convenient sampling. Give a short description of the pretesting mentioning the number of samples, period, and location for it.

Response: We are again thankful to the reviewer. While we had a 5% nonresponse rate in our final survey, we found less than 5% (2 of 50 mothers refused to be involved in the study) as the nonresponse rate during pretesting of our study. The description of the pretest has been given in our revised manuscript. In our main survey, the nonresponse rate was 2%.

Measurement of Knowledge and Practice Score

Give the 15 knowledge-related questions and 13 practice-related questions in the supplementary file. Mention if these questions are your own or if you used any valid tools or questions adopted from the relevant previous studies. Give adequate information regarding the scoring system of the variables, mentioning the highest possible aggregated score and examples of two questions (one for knowledge and one for practice).

Response: We again appreciate the reviewer. The knowledge and practice questions have been added to the supplementary file (Supplementary Table S1 and Table S2). Both knowledge and practice questions were adopted from reviewing the literature and revised according to our selection criteria. The summation scoring technique was used in computation, and their descriptive statistics, including percentiles, were observed. Then, both the knowledge and practice scores were classified according to percentile, which is evident in the existing literature (reference added). The range for the knowledge and practice scores was 1-15 and 1-11, respectively. In the main text, the section has been revised accordingly.

Statistical Analyses

The authors mentioned that they used the Mann-Whitney U test and the Kruskal-Wallis test. However, they did not mention the underlying assumptions of the tests. Moreover, the Results section also shows the χ^2 test but is not mentioned in the Methods section. Furthermore, the last line of the Results of the abstract shows the Pearson correlation coefficient, but nothing is mentioned in the Methods or Results section of the entire manuscript.

Response: We apologize for the mistake. Necessary assumptions were checked before performing statistical analysis. The Statistical Analysis section has been revised and mentions the χ^2 test and Pearson correlation coefficient. All the necessary corrections raised by the editor and reviewers have been addressed.

Results

Table 1

It is confusing as the text description of Table 1 and the title of Table 1 are different. It is recommended to use two separate tables: one for socioeconomic variables and another for the frequency distribution of the knowledge level among socioeconomic variables. Mention the knowledge- and practice-related raw scores first and then the cross-tab results. There is a major mistake in the results of Tables 1 and 2. The frequency distribution for educational status, occupation, family type, number of family members, and monthly income in Tables 1 and 2 are the same. However, the P values are different. How is this possible? Please check the results.

Response: Please accept our apology for the error that happened unconsciously. The frequency distribution for educational status, occupation, family type, number of family members, and monthly income in Tables 1 and 2 has been rechecked and revised. In addition, Table 1 has been separated into two tables (Tables 1 and 2) and presented accordingly.

Discussion

It is confusing whether the practice was for the children or how a mother takes care of their children's dental health. Mention the implications of your findings rather than just comparing the findings with previous studies. State the limitation of the study, especially the bias regarding convenient sampling. Provide a section on the public health significance of the study findings in Bangladesh.

Response: We sincerely appreciate the reviewer for these comments. The Discussion of the manuscript has been revised accordingly. The limitations have been revised in the Discussion section.

Conclusion

The Conclusion section of the study is poorly written and not focused on the findings of the study. Revise the Conclusion section to highlight your study findings.

Response: Thank you again. The Conclusion of the manuscript has been revised accordingly.

Reviewer AJ [3]

Specific Comments

There were a lot of grammatical issues and typographical errors. The manuscript needs to be edited for grammar and syntax. It is also obvious that the manuscript was not proofread adequately.

Major Comments

Abstract

- *A word is missing in the first sentence. Authors should proofread the manuscript.*
- *Keywords: Dhaka is a more appropriate keyword than Bangladesh.*
- *Under the Results in the abstract, respondents should be referred to as such and not as samples.*

Introduction

- *The global prevalence of oral diseases was stated, but authors did not capture the prevalence in the study area/country and so have not shown that oral disease is a problem. Even the global prevalence that was stated was only that of dental caries among the seven conditions that make up oral diseases as stated by the authors.*
- *The objective stated here (last sentence) comes off like the authors are assessing the knowledge and practices of oral hygiene with regard to themselves and not their children as stated in the topic.*

Methods

- *Was it permission that was given by the institutional review board or an ethical clearance?*
- *This section is quite disorganized. There is a logical flow expected in this section.*
- *Why was a nonprobability sampling technique (convenient sampling) used for this study? The sampling technique was not explained at all. This will make replicating this study difficult.*
- *I have an issue with the scoring system and the grading. Is there a reference for it? I particularly have an issue with "moderately average." It is not a standard term.*
- *The exclusion criteria are not the opposite of the inclusion criteria as stated by the authors. Exclusion criteria are those already included in the study but that are ineligible for one reason or the other.*

Results

- *In the text above Table 1, authors wrote that most respondents (39.3%) had a monthly family income of "21,000 - 40,000 taka per month." This figure (39.3%) is just over one-third of the respondents and not a majority.*
- *Table 1: What is the meaning of graduation and above? Is it graduated secondary school or graduated college?*
- *"Respectively" should be added at the end of the following sentence. "Out of 400 mothers, more than 90% knew the importance of brushing teeth while 82.3% and 80.8% of them knew the recommended frequency and appropriate time for brushing teeth."*

Discussion

- *The second sentence: the study is not evaluating parent's knowledge and practices but that of mothers.*
- *Grammatical errors and missing words*

Reference List

- *Some of the references were not cited correctly. Authors should adhere to the Vancouver referencing style.*

Round 2 Review

Reviewer BZ

The authors impressively amended the initial version of the manuscript based on the reviewers' comments. However, several issues remain unaddressed.

1. The authors should include the city in the title of the study. You can revise the title to "Knowledge and practices towards oral hygiene of children aged 5 - 9 years old: a cross-sectional study among mothers visited tertiary level hospitals in Dhaka, Bangladesh."

Response: Thanks for this suggestion. We revised the title of the manuscript accordingly as "Knowledge and practices towards oral hygiene of children aged 5 - 9 years old: a cross-sectional study among mothers visited tertiary level hospitals in Dhaka, Bangladesh."

2. Use the full form when it appears first and then use the abbreviation afterward. For example, "KP" in the abstract.

Response: Thanks again for this suggestion. We revised the title of the manuscript accordingly.

3. Please mention this statistical test in the Methods section of the abstract. You did not mention the χ^2 test and Pearson correlation.

Response: Revised the Methods section of the manuscript accordingly as "Statistical analysis including the χ^2 test and Pearson correlation test were performed. The Mann–Whitney *U* test and Kruskal–Wallis one-way ANOVA test were performed to show average knowledge and practice variations among different socio-demographics groups."

4. It is recommended to make the recommendation simple and easy to understand for the readers. Avoid duplication of the same term.

Response: Revised the Recommendation section accordingly.

5. In the sample size calculation, you used P=.58 and P=.57. Please clarify why you used those prevalences. Cite the relevant study here.

Response: The Sample Size Calculation section has been revised accordingly as "A convenient sampling technique was followed for this study. During literature search, no study was found that assessed knowledge and practice towards children's oral hygiene among Bangladeshi mothers. But, a very few studies found in other country with similar socio-demography (eg, India). Mohandas et al, 2021 in his study entitled 'Knowledge and practice of rural mothers on oral hygiene for

children' showed the prevalence of knowledge and practice were 58% and 57% respectively [4]. The sample size was calculated using the below equation.

"n =(z^2 pq)/d^2 …… (1)

"the sample size for the mother's knowledge when *P*=.58 was

"n=(〖1.96〗 ^2×0.58 × (1 - 0.58))/ 〖0.05〗 ^2=375

"Similarly, the sample size for mother's practice level when *P*=.57 was

"n=(〖1.96〗 ^2×0.57 × (1 - 0.57))/ 〖0.05〗 ^2=377

"Therefore, we initially chose a maximum of 377 as the required sample size. Considering a maximum 5% non-response rate (based on pre-testing), we rounded up this figure and selected 400 as the approximate sample size in the study."

6. Before the heading for the sociodemographic variables in the Methods section, you mention outcome measures. However, the sociodemographic variables are not your outcome variables according to your objectives. You can remove the term outcome measures from here.

Response: The heading "Outcome measure" has been removed from the revised manuscript.

7. You mentioned that you used 13 questions for the assessment of practices. Thus, according to your scoring approach, there should be a score of 1-13, but here, it is 1-11.

Response: Thank you again. We revised the error. The change is "The range for knowledge and practice score was 1 to 15, and 1 to 13 respectively."

8. Please mention the name of the software and version you used for the statistical analysis.

Response: Thank you again. We added the statistical software name with the version as "All the data management and statistical analyses were carried out through IBM SPSS Statistics 25.0."

9. Revise the sentence before Table 1. You can make it two sentences. One for family income and another for occupation.

Response: We revised the sentence accordingly as "Majority of the respondents (39.3%) had the monthly family income of 21000 - 40000 ($206.19-$392.73) Taka per month. About 13.3% mothers were involved in any paid worked activities (Table 1)."

10. There is no chi-square–related data in Table 1. Please remove the footnotes from Table 1.

Response: Removed the errors.

11. In Figure 1, it is recommended to keep the values to one decimal point for 1a and 1b.

Response: Thank you for this suggestion. We removed Figures 1c and 1d in our revised manuscript.

12. Please revise the sentence before Table 3 to give a clear meaning.

Response: We revised the sentence accordingly as "The educational status (*P*=.002) and income (*P*=.044) were significantly associated with mothers' oral hygiene practices (Table 3)."

13. You can remove the percentage symbol from the value and give it in the vertical axis title.

Response: Removed accordingly.

14. Please give the correlation results in the main manuscript or as a supplementary table.

Response: The correlation results have been given as the supplementary result. Please see Supplementary Result S6.

15. The authors overlooked the association of knowledge and practice with income and family size. Please give more details on those two points in the Discussion section.

Response: The variable family income has been addressed in the Discussion. Please see page 17 (before the Strengths and Limitation section). Family income has been discussed briefly in the Principal Findings section.

Conflicts of Interest

None declared.

References

1. Islam MH. Peer review of "Mothers' Knowledge of and Practices Toward Oral Hygiene of Children Aged 5-9 Years in Bangladesh: Cross-Sectional Study". JMIRx Med 2025;6:e70144. [doi: 10.2196/70144]
2. Tamannur T, Das SK, Nesa A, et al. Mothers' knowledge of and practices toward oral hygiene of children aged 5-9 years in Bangladesh: cross-sectional study. JMIRx Med 2025;6:e59379. [doi: 10.2196/59379]
3. Nwankwo B. Peer review of "Mothers' Knowledge of and Practices Toward Oral Hygiene of Children Aged 5-9 Years in Bangladesh: Cross-Sectional Study". JMIRx Med 2025;6:e70142. [doi: 10.2196/70142]
4. Mohandass B, Chaudhary H, Pal GK, Kaur S. Knowledge and practice of rural mothers on oral hygiene for children. Indian J Continuing Nurs Education 2021;22(1):39-43. [doi: 10.4103/IJCN.IJCN_7_20]

Edited by T Leung; submitted 16.12.24; this is a non–peer-reviewed article;accepted 16.12.24; published 03.02.25.

Please cite as:
Tamannur T, Das SK, Nesa A, Nahar F, Nowshin N, Binty TH, Shakil SA, Kundu SK, Siddik MAB, Rafsun SM, Habiba U, Farhana Z, Sultana H, Kamil AA, Rahman MM
Authors' Response to Peer Reviews of "Mothers' Knowledge of and Practices Toward Oral Hygiene of Children Aged 5-9 Years in Bangladesh: Cross-Sectional Study"
JMIRx Med 2025;6:e70145
URL: https://xmed.jmir.org/2025/1/e70145
doi:10.2196/70145

Author's Response to Peer Reviews of "Converting Organic Municipal Solid Waste Into Volatile Fatty Acids and Biogas: Experimental Pilot and Batch Studies With Statistical Analysis"

Hojjat Borhany, MSc

Faculty of Environmental Science, Department of Environmental Science, Informatic, and Statistics, University of Ca' Foscari Venice, Mestre (VE), Italy

Corresponding Author:
Hojjat Borhany, MSc
Faculty of Environmental Science, Department of Environmental Science, Informatic, and Statistics, University of Ca' Foscari Venice, Mestre (VE), Italy

Related Articles:

Companion article: https://www.biorxiv.org/content/10.1101/2023.06.21.545938v2

Companion article: https://med.jmirx.org/2025/1/e69895

Companion article: https://med.jmirx.org/2025/1/e69896

Companion article: https://med.jmirx.org/2025/1/e50458

(JMIRx Med 2025;6:e69894) doi:10.2196/69894

KEYWORDS

multistep fermentation; specific methane production; anaerobic digestion; kinetics study; biochar; first-order; modified Gompertz; mass balance; waste management; environment sustainability

This is the author's response to peer-review reports of "Converting Organic Municipal Solid Waste Into Volatile Fatty Acids and Biogas: Experimental Pilot and Batch Studies With Statistical Analysis."

Round 1 Review

Anonymous [1]

The present manuscript [2] deals with the study of the valorization of organic fractions of municipal solid waste through the production of volatile fatty acids (VFAs) and biogas. The article is interesting; in my opinion, it should be revised.

Comments

1. The presentation of the manuscript is very poor; the figures are not in the same format.

Response: The remaining figures, which included the box plots of VFA concentration, VFA/soluble chemical oxygen demand (SCOD) ratio, scheme of line, VFA and SCOD concentration, VFA weight ratio distribution, capital cost and yearly income, and biomethane content, were kept and reformulated to have the same shape. The figures outlining the kinetics study were deleted.

2. Some of the recent works should be discussed and cited in the Introduction section: [3-7].

Response: Some of the recent relevant works and studies were discussed and cited in the Introduction section as follows:

- Inyang M, Gao B, Pullammanappallil P, Ding W, Zimmerman AR. Biochar from anaerobically digested sugarcane bagasse. *Bioresour Technol*. Nov 2010;101(22):8868-8872. [doi: 10.1016/j.biortech.2010.06.088] [Medline: 20634061]
- Jung S, Shetti NP, Reddy KR, et al. Synthesis of different biofuels from livestock waste materials and their potential as sustainable feedstocks – a review. *Energy Conversion Manage*. May 15, 2021;236:114038. [doi: 10.1016/j.enconman. 2021.114038]
- Sampath P, Brijesh, Reddy KR, et al. Biohydrogen production from organic waste – a review. *Chem Eng Technol*. Jul 2020;43(7):1240-1248. [doi: 10.1002/ceat.201900400]
- Algahashm S, Qian S, Hua Y, et al. Properties of biochar from anaerobically digested food waste and its potential use in phosphorus recovery and soil amendment. *Sustainability*. Dec 10, 2018;10(12):4692. [doi:10.3390/su10124692]

3. The novelty of the work should be highlighted.

Response: We noted at the end of the Introduction and at the beginning of the Discussion that this study is novel in that it presents a strong framework for evaluating a proposal for the

financial and technical valorization of organic municipal solid waste using statistical analysis, process kinetics, mass balance, and experimental testing. Furthermore, as compared to single-step anaerobic digestion, our data showed a notably high improvement in profitability and a corresponding decrease in the payback period. In order to further close the cycle circuit and prolong the product life, we also proposed the integration of two potential future units.

4. Full stops should be removed from all subheadings.

Response: They are all removed.

5. The Results and Discussion should be written in detail with proper subheadings.

Response: The Results section was rewritten and divided into subheadings to mirror their counterparts in the Methods, and the Discussion section has the added subheadings Principal Results, Comparison With Previous Works, and Conclusion and Limitations according to the required information in the guidelines of JMIR Publications.

6. There are some typo errors; they should be rectified.

Response: They were corrected.

Reviewer GA [8]

General Comments

Generally, the manuscript should be strictly improved in English language writing and corrected for all grammatical errors throughout the whole manuscript. The author has to use a uniform style of the English language, either American or British English. Further English assistance is particularly required. Many missing articles and a lot of grammatical and punctuation errors must be corrected in the manuscript as in the corrected abstract.

Response: The abstract was prepared in an organized format and corrected for its language. We also employed English assistance. The manuscript's English was improved, and its style was harmonized with American English.

Specific Comments

This paper shows an important aspect of multiple fermentation steps for the complete utilization of municipal solid waste and conversion to useful products, which is highly recommended for circular economic sustainability worldwide. However, it needs some major revision and arrangement to allow for a better presentation of this valuable work.

Major Comments

Title

1. "Valorization of Organic Fraction of Municipal Solid Waste Through Production of Volatile Fatty Acids (VFAs) and Biogas" is a long title that should be shortened to be more concise with no abbreviations—more indicative. Suggested title: "Valorization of Organic Municipal Solid Waste for Volatile Fatty Acids and Biogas Production."

Response: It was adopted according to the guidelines for the descriptive title of the original paper: "Issue/Intervention in Demographic/Disease/Condition: Method/Study Design"; "Conversion of Organic Municipal Solid Waste to Volatile Fatty Acids and Biogas: Experimental Pilot and Batch Studies with Statistical Analysis."

Abstract Section

2. General language; it must be more concise and specific.

Response: I did search for all the general language in the manuscript and tried to provide concise information on the matter.

3. Please clearly mention the take-home message and the main findings of the research.

Response: The research's primary conclusions include the development of a reliable technique for evaluating the recovery proposal for the conversion of organic solid waste into valuable products and assessing both its technical and financial viability. Furthermore, our proposal outperforms the conventional approaches in terms of economics.

4. The abstract is too long and lacks the main methodology and main experimental techniques that were carried out in this work. The author may add some hints about the main methods used before mentioning the main results.

Response: Subheadings for the background, objective, method, findings, and conclusion were added to the revised abstract. There are fewer words in the abstract overall than the 450-word limit. Additionally, some pointers regarding experimental techniques such as gas chromatography are provided, along with the kind of statistical test used to verify the significance and efficacy of the suggested process amendment. We also mentioned the use of mass flow models for the process's economic evaluation and the various kinetics models that can be used to describe biogas production.

Manuscript

5. Keywords: Words must be modified to be more informative and representative of the research interest and differ from the word in the manuscript title. Maybe add "Multi Step of Fermentation Process" or "Waste Management and Environment Sustainability."

Response: We updated the keywords to include "Multistep Fermentation," "Environment Sustainability," "Waste management," "Specific Methane Production," "Anaerobic Digestion," "Kinetics Study," "Biochar," "First-Order," "Modified Gompertz," and "Mass Balance."

6. Arrangement of the experimental work in the manuscript may be needed in the Results and Discussion accordingly.

Response: It was completed in a way that would make it easier for specialists in the field to follow the stages, and a Discussion section was included to compare the findings with earlier research, highlight the key conclusions, and clarify the research's limitations.

7. There is a lack of figures to describe the main parameter optimization steps well. Please reformulate to describe some data using figures with error bars.

Response: Our optimization procedure focused on reducing the payback period by decreasing the cost and increasing the profit from bioproducts. This was achieved through pilot tests for examining the effectiveness of the hydraulic retention time (HRT) manipulation and pretreatment in increasing the VFA yield and the integration of our process knowledge of using the fine-tuned feedstock/inoculum ratio as well as biochar addition to obtain the biogas in a cost-effective process. Detailed information and calculations regarding the mass flow analysis are available in the supplementary documents in the Excel spreadsheet named"Mass Balance.". For figures, we provided the VFA concentrations and distribution for two HRTs and a *t* test to confirm the significance of the results. Further, for biogas production, we provide results from a kinetics study showing an 8-fold increase in the hydrolysis rate and a 100% decrease in the lag phase. This brought about a small anaerobic digester working at a high organic loading rate, leading to a reasonably priced process.

8. The SD and table footnotes with the number of replicates should be noted underneath all of the given tables.

Response: For all data that was accompanied by an SD, the number of replicates was reported beneath all the given tables.

9. A mechanistic in-detail discussion is required, not just comparing your results with the previous work; justify better.

Response: The comparisons of results from similar studies were done mechanistically and in detail.

For example:

- "Because of the extra pretreatment unit in our study, our VFA yield was significantly higher than the study by valentino et al "
- "The higher hydrolysis rate was due to the destruction of the solids structure caused by bacterial enzymes and a hot alkaline solution. Additionally, we provided a higher active biomass per feedstock using a fine-tuned FS/IN ratio of 0.3 (VS basis), which was noticeably lower than the quantities (1 and 0.5) reported in similar studies "
- "due to the added fresh WS with higher digestible content and better nutrient balance than the fermented solids, the SMP value by valentino et al was higher."
- "The higher practicability than the 2 steps of bioethanol and biogas production as a result of sterilization and high bioethanol concentration requirements."
- "Our proposal is more favorable since it does not limit the VFA weight ratio distribution and does shifts the recovery route toward higher market-valued products like VFAs than single step AF + AD by Papa et al"

10. In research articles, do not include any table comparing literature results; the author can discuss the main findings in the text itself, as in Table 5.

Response: All the data in the tables comparing results were deleted, and we discussed them in the text.

11. The Conclusions section is missing in the manuscript to summarize and point out the novelty and the main findings from the research.

Response: The Conclusion was included in the manuscript and presents the main findings as follows: "To conclude, we presented a robust framework to assess a proposal for the valorization of organic waste through experimental tests, statistical analysis, and process kinetics, along with mass and energy flow analysis. The findings support considerably higher profitability and, as a result, a shorter payback period for multistep reclamation than the current single anaerobic digestion. Further, our results encourage the circular economy perspective on the conversion of OMSW into biogas and VFAs, with the pros of fewer residual solids due to reusing them in a pyrolysis line."

12. Generally speaking, in academic writing, (1) abstracts do not include abbreviations, (2) avoid articles in the title (the, a, an), and (3) avoid keywords that exist in the title.

Response: (1) Based on JMIR House Style and Guidelines, the usage of abbreviations and acronyms in the abstract section is not forbidden. Further, all author-invented abbreviations were omitted. We also stop using "AD" as an abbreviation for anaerobic digestion since it may make it ambiguous with "AD" (the reference year). In fact, keeping the number of words in the abstract within the limits is really impossible without using some of them. (2) It was avoided. (3) It was avoided to be as informative as possible.

13. As a rule of thumb, no dots in titles or subtitles as in the Experimental section, Anerobic Pilot Unities, etc.

Response: The dots were removed.

14. Multiple references should be merged, not written separately, as in "29, 30" and "23, 27"; the author may use the merge reference option in reference software.

Response: It was corrected.

15. The author may add numbers for all titles and subtitles accordingly all over the manuscript.

Response: Based on the JMIR guidelines for the author, it is not allowed to use numbering for headings and subheadings.

Minor Comments

16. The author should avoid general and well-known information, and be selective in the recent references used. May add one small paragraph to the Biological Waste Management and Environment Sustainability section.

Response: The small paragraph already discussed the current state of municipal organic waste production and treatment in the European Union. We extended it and incorporated all other information regarding environmental sustainability from some relevant sources suggested by the peer reviewer.

17. The author should clarify the main aim of the work clearly in the last paragraph of the Introduction.

Response: The main aim of this study was an assessment of multistep pretreatment acidogenic fermentation, followed by anaerobic digestion of municipal organic waste in comparison with the existing method of single anaerobic digestion in terms of financial profit and technical feasibility.

18. Do not use our, we, or us in academic writing.

Response: Based on the journal guidelines, there are no issues with using we and us in the article submitted to JMIR Publications; nevertheless, I do my best to avoid overusing these words in my manuscripts.

19. The author may mention novel applications of VFA and biogas. Mention different novel sources of biogas production.

Response: It was already mentioned in the study that biogas and VFA typically were used for energy production and biopolymer synthesis, respectively. Moreover, other sources of biogas typically were from nonbiological processes, which were beyond our scope since we focused on carbon-neutral microbiological processes.

20. The author should mention the gas chromatography type, gas injection rate, column dimensions, and the used carrier gas in the main document.

Response: It was included in the Methods section.

21. The author did not mention that flushing with nitrogen or carbon dioxide took place in anaerobic digestion while feeding reactors and how the anaerobic conditions were maintained; please mention it clearly or add the references used for the methodology.

Response: The anaerobic condition was ensured in bottles just by sealing them after filling without any flushing with nitrogen or carbon dioxide since we had known that the oxygen transfer at the surface of the waste stream was impossible as it contained high total solids and SCOD. This type of procedure was adopted in our lab and has been conducted for years.

22. Organize titles all over the manuscript.

23. Generally, the subtitles are too generic; modify them to be more indicative and precise.

Response: The subtitles were modified to be more indicative and precise.

24. "unless Saturday and Sunday" in line 208 is not important information; the suggested word "daily" is enough.

Response: It was corrected.

25. "Unite": Please correct.

Response: All units are corrected.

26. Remove the grid lines in the figures.

Response: They were removed.

27. The author has to mention the range used for the chemical oxygen demand method, and the original reference should be cited appropriately.

Response: The method for determination of soluble and solid chemical oxygen demand of the waste stream was according to the Standard Methods for Water and Wastewater. We also clearly discussed in the Methods section a proper limit of detection and reference.

28. "As can be seen": This statement is repetitive more than once in the Discussion, in lines 301, 315, and 423.

Response: Line 301 was corrected. Line 315 was corrected to be informative and avoid repetition. Line 423 was rectified in English language, and the repetitive statements were removed.

29. Figure 3 caption: Mesophilic fermentation: Please specify which stage because both of the sequential steps were called mesophilic fermentation in Figure 1.

Response: In fact, Figure 3 depicts the weight ratio distribution from the second step named mesophilic acidogenic fermentation. Surprisingly, the VFA could only be obtained from the second stage. Additionally, we modified the caption to read "VFAs weight ratio distribution for mesophilic acidogenic fermentation" and made a clear reference to Figure 1, which depicts the processes of pretreatment, acidogenic fermentation followed by mesophilic anaerobic digestion . In terms of pH and HRT, the two later procedures differ from one another substantially.

30. What is the rationale for comparing 3 days to 4.5 days for all the used systems; the author may justify why 4.5 days is better to complete with this HRT in the rest of the experiments or describe the one variable at a time optimization method that is used to determine the significant factors and the insignificant one; mention them clearly. Also, use in the Discussion the terms "significant" and "insignificant" according to the obtained P value.

Response: The values for the two HRTs to increase the VFA concentration in the outlet were selected based on our experience and process knowledge. According to this information, exceeding the HRT value of more than 3 - 5 days can bring the process into an anaerobic digestion step. As a result, the VFAs with high-added value markets are converted to biogas. Hence, the two HRTs of 3 days and 4.5 days were tried in the pilot test, knowing that the VFA concentration would either increase or decrease linearly in this local region of operation.

31. The author has to mention tables and figures in the text in their appropriate place.

Response: They were mentioned where they were referred to.

References

1. Anonymous. Peer review of "Converting Organic Municipal Solid Waste Into Volatile Fatty Acids and Biogas: Experimental Pilot and Batch Studies With Statistical Analysis". JMIRx Med 2025;6:e69895. [doi: 10.2196/69895]
2. Borhany H. Converting organic municipal solid waste into volatile fatty acids and biogas: experimental pilot and batch studies with statistical analysis. JMIRx Med 2025;6:e50458. [doi: 10.2196/50458]
3. Jung S, Shetti NP, Reddy KR, et al. Synthesis of different biofuels from livestock waste materials and their potential as sustainable feedstocks – a review. Energy Convers Manage 2021 May 15;236:114038. [doi: 10.1016/j.enconman.2021.114038]

4. Srivastava RK, Shetti NP, Reddy KR, Aminabhavi TM. Sustainable energy from waste organic matters via efficient microbial processes. Sci Total Environ 2020 Jun 20;722:137927. [doi: 10.1016/j.scitotenv.2020.137927] [Medline: 32208271]
5. Sampath P, Brijesh, Reddy KR, et al. Biohydrogen production from organic waste – a review. Chem Eng Technol 2020 Jul;43(7):1240-1248. [doi: 10.1002/ceat.201900400]
6. Velvizhi G, Goswami C, Shetti NP, Ahmad E, Kishore Pant K, Aminabhavi TM. Valorisation of lignocellulosic biomass to value-added products: paving the pathway towards low-carbon footprint. Fuel (Lond) 2022 Apr 1;313:122678. [doi: 10.1016/j.fuel.2021.122678]
7. Monga D, Shetti NP, Basu S, et al. Engineered biochar: a way forward to environmental remediation. Fuel (Lond) 2022 Mar 1;311:122510. [doi: 10.1016/j.fuel.2021.122510]
8. Elsalamony D. Peer review of "Converting Organic Municipal Solid Waste Into Volatile Fatty Acids and Biogas: Experimental Pilot and Batch Studies With Statistical Analysis". JMIRx Med 2025;6:e69896. [doi: 10.2196/69896]

Abbreviations

HRT: hydraulic retention time
SCOD: soluble chemical oxygen demand
VFA: volatile fatty acid

Edited by T Leung; submitted 10.12.24; this is a non–peer-reviewed article;accepted 10.12.24; published 04.02.25.

Please cite as:
Borhany H
Author's Response to Peer Reviews of "Converting Organic Municipal Solid Waste Into Volatile Fatty Acids and Biogas: Experimental Pilot and Batch Studies With Statistical Analysis"
JMIRx Med 2025;6:e69894
URL: https://xmed.jmir.org/2025/1/e69894
doi:10.2196/69894

Authors' Response to Peer Reviews of "Determinants of Periodic Health Examination Uptake: Insights From a Jordanian Cross-Sectional Study"

Abdul Aziz Tayoun, MPH

School of Medicine, Department of Family and Community Medicine, Jordan University, Qween Rania Street, Amman, Jordan

Corresponding Author:
Abdul Aziz Tayoun, MPH
School of Medicine, Department of Family and Community Medicine, Jordan University, Qween Rania Street, Amman, Jordan

Related Articles:

Companion article: https://www.medrxiv.org/content/10.1101/2024.02.03.24302286v1

Companion article: https://med.jmirx.org/2025/1/e71529

Companion article: https://med.jmirx.org/2025/1/e71531

Companion article: https://med.jmirx.org/2025/1/e57597

(JMIRx Med 2025;6:e71528) doi:10.2196/71528

KEYWORDS

periodic health examination; PHE; preventive health services; routine health checkups; Jordan; cross-sectional study

This is the author's response to peer-review reports for "Determinants of Periodic Health Examination Uptake: Insights From a Jordanian Cross-Sectional Study."

Round 1 Review

Anonymous [1]

The following items were noted in this paper [2].

- *Periodic health examination (PHE) uptake: Only 27.1% of participants underwent a PHE in the last 2 years.*
- *Predictors: Significant predictors include recent visits to a primary health care facility, monthly income, and knowledge about PHEs and preventive health measures.*
- *Nonsignificant factors: Gender, marital status, smoking status, and BMI did not show a significant association with PHE uptake.*

Strengths

- *Comprehensive analysis: The study employs a robust methodology, combining descriptive, inferential, and multivariate statistical techniques to provide a thorough understanding of PHE uptake.*
- *Significant predictors identified: Key factors influencing PHE uptake were identified, offering valuable insights for health care providers and policy makers.*
- *First of its kind in Jordan: This study fills a gap in existing knowledge by being the first to investigate PHE uptake in Jordan.*

Negative Points and Areas for Improvement

Cross-Sectional Design

- *Limitation: The study's design limits the ability to establish causality.*
- *Improvement: Future research could benefit from a longitudinal approach to better establish causal relationships between the identified predictors and PHE uptake.*

Response: We acknowledge the limitation of the cross-sectional design in establishing causality and have highlighted this in the Discussion section, suggesting future longitudinal studies.

Convenience Sampling

- *Limitation: This method may introduce selection bias, and the online survey format may lead to measurement bias.*
- *Improvement: Employing a more randomized and stratified sampling method could enhance the representativeness and validity of the findings.*

Response: We have clarified the rationale for using convenience sampling due to resource constraints and have suggested more randomized methods for future studies.

Limited Generalizability

- *Limitation: Results may not be generalizable to populations outside of Jordan or those not included in the sample.*
- *Improvement: Expanding the study to include diverse populations and different geographic regions would provide a more comprehensive understanding of PHE uptake.*

Response: We understand the concern regarding generalizability. However, as the study aimed to estimate PHE uptake and its determinants specifically in Jordan, the focus on this population was intentional. For future research, we recommend conducting multinational studies, particularly in Arab countries, or performing systematic reviews or meta-analyses to obtain results that can be generalized beyond Jordan.

Survey Instrument

- *Limitation: The questionnaire's comprehensiveness and relevance to the Jordanian context might not have been fully ensured.*
- *Improvement: Pretesting the survey with a larger and more varied group, followed by adjustments based on feedback, could improve its applicability and accuracy.*

Response: We have taken steps to improve the relevance and comprehensiveness of the questionnaire by pretesting it and incorporating feedback.

Behavioral Factors

- *Limitation: The study did not find a relationship between behavioral factors and PHE uptake, which contradicts findings in other contexts.*
- *Improvement: A more detailed investigation into cultural and societal influences on health behaviors in Jordan is needed to clarify these results.*

Response: We agree that further investigation into cultural and societal influences on health behaviors in Jordan is needed and have discussed this in the manuscript.

English Language and Clarity

- *Limitation: The manuscript contains some grammatical errors and awkward phrasings, which can detract from its readability.*
- *Improvement: A thorough review and editing for language and clarity by a native English speaker or professional editor would enhance the manuscript's quality.*

Response: The manuscript has undergone a thorough review and editing process to enhance its readability and clarity.

Thank you for these excellent comments. We have thoroughly reviewed and integrated your suggestions into the main manuscript.

Reviewer AV [3]

Specific Comments

Major Comments

1. In this manuscript, write in detail about the data collection procedure.

Response: The data collection process was reviewed in detail. Please refer to the Methodology section and note that the questionnaire has been added as an appendix (see Multimedia Appendix 1).

2. Why was a convenience sampling technique employed?

Response: A convenience sampling technique was employed due to resource constraints, as the study was not funded and was conducted by a single author. This has been mentioned in the Methodology section.

3. "All collected data are treated with strict confidentiality." Some language corrections are required.

Response: We have rephrased the Ethical Consideration section to improve clarity and accuracy.

Minor Comments

There are a lot of formatting issues; many things seem copied and pasted.

Response: We have addressed the formatting issues to ensure consistency and clarity throughout the document.

Conflicts of Interest

None declared.

References

1. Anonymous. Peer review of "Determinants of Periodic Health Examination Uptake: Insights From a Jordanian Cross-Sectional Study". JMIRx Med 2025;6:e71531. [doi: 10.2196/71531]
2. Tayoun AA. Determinants of periodic health examination uptake: insights from a Jordanian cross-sectional study. JMIRx Med 2025;6:e57597. [doi: 10.2196/57597]
3. Ahmed A. Peer review of "Determinants of Periodic Health Examination Uptake: Insights From a Jordanian Cross-Sectional Study". JMIRx Med 2025;6:e71529. [doi: 10.2196/71529]

Abbreviations

PHE: periodic health examination

Edited by T Leung; submitted 20.01.25; this is a non–peer-reviewed article;accepted 20.01.25; published 05.02.25.

Please cite as:
Tayoun AA
Authors' Response to Peer Reviews of "Determinants of Periodic Health Examination Uptake: Insights From a Jordanian Cross-Sectional Study"
JMIRx Med 2025;6:e71528
URL: https://xmed.jmir.org/2025/1/e71528
doi:10.2196/71528

Authors' Response to Peer Reviews of "Checklist Approach to Developing and Implementing AI in Clinical Settings: Instrument Development Study"

Ayomide Owoyemi[1], MScPH, MD, PhD; Joanne Osuchukwu[2], MD; Megan E Salwei[3], BSc, MSc, PhD; Andrew Boyd[1], BSc, MD

[1]Department of Biomedical and Health Informatics, University of Illinois Chicago, 1919 W Taylor, Chicago, IL, United States

[2]College of Medicine, University of Cincinnati, Cincinnati, OH, United States

[3]Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, TN, United States

Corresponding Author:
Ayomide Owoyemi, MScPH, MD, PhD
Department of Biomedical and Health Informatics, University of Illinois Chicago, 1919 W Taylor, Chicago, IL, United States

Related Articles:

Companion article: https://www.medrxiv.org/content/10.1101/2024.08.08.24311701v1

Companion article: https://med.jmirx.org/2025/1/e69869

Companion article: https://med.jmirx.org/2025/1/e70058

Companion article: https://med.jmirx.org/2025/1/e69593

Companion article: https://med.jmirx.org/2025/1/e69594

Companion article: https://med.jmirx.org/2025/1/e69870

Companion article: https://med.jmirx.org/2025/1/e69595

Companion article: https://med.jmirx.org/2025/1/e65565

(JMIRx Med 2025;6:e69537) doi:10.2196/69537

KEYWORDS

artificial intelligence; machine learning; algorithm; analytics; AI deployment; human-AI interaction; AI integration; checklist; clinical workflow; clinical setting; literature review

This is the authors' response to peer-review reports for "Checklist Approach to Developing and Implementing AI in Clinical Settings: Instrument Development Study."

Round 1 Review

Anonymous [1]

The paper [2] presents the Clinical Artificial Intelligence (AI) Sociotechnical Framework (CASoF), a structured approach to guide the planning, design, development, and implementation of AI systems in health care settings. The framework is designed to address the gap between technical performance and sociotechnical factors that are essential for successful AI deployment in clinical environments.

The authors conducted a literature synthesis and a modified Delphi study involving global health care professionals to develop and refine the CASoF checklist. The checklist emphasizes the importance of considering the value proposition, data integrity, human-AI interaction, technical architecture, organizational culture, and ongoing support and monitoring, to ensure that AI tools are not only technologically sound but also practically viable and socially adaptable within clinical settings.

The study found that the successful integration of AI in health care depends on a balanced focus on both technological advancements and the sociotechnical environment of clinical settings. The CASoF represents a step forward in bridging this divide, offering a holistic approach to AI deployment that is mindful of the complexities of health care systems. The checklist aims to facilitate the development of AI tools that are effective, user-friendly, and seamlessly integrated into clinical workflows, ultimately enhancing patient care and health care outcomes.

The authors acknowledge some limitations of the study, such as the need for continuous refinement of the CASoF through iterative feedback and broader engagement with more stakeholders. Future research should aim to include an even wider array of perspectives, particularly from underrepresented regions and specialties, to enhance the framework's comprehensiveness and applicability.

Overall, the paper provides a valuable contribution to the field of AI in health care by offering a practical and comprehensive approach to the development and implementation of AI systems in clinical settings.

Reviewer AE [3]

General Comments

This paper presents the Clinical Artificial Intelligence (AI) Sociotechnical Framework (CASoF), a checklist intended to support the development and implementation of AI systems in health care settings. The framework is built on a comprehensive literature review and a modified Delphi study involving health care professionals globally. The manuscript addresses a significant gap in the integration of AI by emphasizing the importance of sociotechnical considerations alongside technical aspects.

Specific Comments

Major Comments

1. Clarity and structure: The manuscript could benefit from clearer explanations, particularly in the methodology section. The description of the Delphi study and literature synthesis is dense and may be difficult for readers to follow. Consider breaking down complex sentences and using more straightforward language.

Response: Thank you for this; we have addressed and improved on the clarity and description of the methodology section as requested.

2. Methodological rigor: The manuscript lacks details on the selection process for Delphi panelists and the exact methods used for data analysis. Transparency in these areas would significantly strengthen the paper. Additionally, clarify how the Delphi method was modified and the rationale behind these modifications.

Response: We have addressed the selection process and what the modification of the Delphi process involves.

3. Literature review and contextualization: The discussion section could benefit from a more critical comparison between the CASoF and existing frameworks. While the manuscript mentions other frameworks, it does not fully explore their limitations or how the CASoF overcomes these challenges. Expanding this discussion would provide a stronger justification for the CASoF's novelty and utility.

Response: We have added important comparisons with other existing frameworks/checklist and what utility the Clinical Artificial Intelligence (AI) Sociotechnical Framework (CASoF) has over them.

4. Checklist practicality: While the checklist is comprehensive, its length and complexity may hinder practical adoption. Consider providing a condensed version for quick reference and include examples of how the checklist can be applied in real-world scenarios.

Response: The application of the checklist in a real-world scenario has been highlighted. We appreciate the suggestion on providing a condensed version; however, we will retain the checklist in its present state and level. We created an online version to make the application easier [4].

5. Ethical considerations and bias mitigation: The manuscript discusses ethical considerations but lacks specific strategies for addressing these issues within the CASoF. Expanding this discussion would enhance the manuscript's comprehensiveness.

Response: The checklist highlights specific questions that addresses ethical considerations; this has also been better highlighted in the manuscript.

Minor Comments

6. Typographical and grammatical errors: There are minor typographical and grammatical errors throughout the manuscript that should be corrected. For instance, phrases like "revised and edited" could be simplified to "revised" for conciseness.

Response: Thanks for this comment; this has been corrected.

7. Tables and figures formatting: The tables summarizing the Delphi study results are helpful but could be more effectively formatted. Using shading or color coding to distinguish between different stages or domains would improve clarity and ease of interpretation.

Response: Thanks, this is well noted. The final formatting would be more of a decision of the publisher.

8. Recent references: Some references in the manuscript are relatively old, which is less ideal for a rapidly evolving field like AI. Where possible, the manuscript should incorporate more recent literature to support its claims and demonstrate the ongoing relevance of the topic.

Response: The references for the articles were selected based on their relevance to the topic.

Reviewer AP [5]

General Comments

This paper...is a very cohesive approach to establishing a framework for the implementation of artificial intelligence (AI).

Specific Comments

Major Comments

1. Ideally there should be information on the demographics of the expert panel.

2. Please add comments regarding equitable access for these technologies.

Response: We did not collect demographic data for the panelists except their professions.

Reviewer BH [6]

General Comments

Using artificial intelligence (AI) to add social and domain-specific steps to clinical trials is innovative. My only comment is whether the number of stages or the checklist changes if the shortlisted panelists change.

Response: This change does not affect the number of changes. The process ends when consensus is reached.

Specific Comments

Major Comments

1. I am unsure if having 38 (expert) panelists is good enough to have a robust framework.

Response: Nasa et al [7] highlighted that a panel of 30 - 50 is considered optimum for a Delphi study.

Anonymous [8]

General Comments

This paper construct a checklist to support the development and implementation of artificial intelligence (AI) in clinical settings. I only have some minor comments.

Minor Comments

1. Comparison with existing checklists: Please add a comparison with some of the existing checklists.

Response: Thank you for this; we have added the necessary comparisons.

2. Inconsistency in the number of studies: The authors initially stated that they included 20 studies, but later mentioned 23. Please clarify.

Response: This has been corrected. There were 19 studies, 3 were excluded, and then 4 were added, which gives a final total of 20.

3. Appendix visibility: The appendix is not visible.

Response: This has been corrected.

4. Abbreviation consistency: The abbreviation "IQR" appears multiple times. Please ensure it is clearly defined and used consistently.

Response: This has been corrected. Thanks.

Anonymous [9]

This paper introduces the Clinical Artificial Intelligence (AI) Sociotechnical Framework (CASoF), a checklist developed through a literature synthesis and refined by a Modified Delphi study. It aims to guide the development and implementation of AI in clinical settings, focusing on the integration of both technological performance and sociotechnical factors. The framework addresses gaps in existing frameworks by emphasizing not only technical specifications but also the broader sociotechnical dynamics essential for successful AI deployment in health care.

New approaches to reporting AI in clinical settings are crucial as AI becomes more integrated into clinical practice. However, the paper needs to address the "black box" dilemma more thoroughly. This refers to the opaque nature of AI algorithms, where the decision-making process is not easily interpretable by clinicians, leading to trust and transparency issues. Additionally, while the CASoF checklist is a valuable tool, it would benefit from a more detailed comparison to established frameworks like TRIPOD (Transparent Reporting of a Multivariable Prediction Model for individual Prognosis or Diagnosis), which has been widely used in developing and validating clinical prediction models. Discussing how the CASoF complements or improves upon TRIPOD would strengthen the paper's contributions.

I suggest adding a paragraph discussing the potential roles of AI when integrated into hospital electronic health record (EHR) systems. AI could be used for the development of advanced diagnostic and prognostic tools by analyzing real-time patient data. Integration with EHRs could enhance decision-making, providing predictive analytics at the point of care and improving patient outcomes. This would help explore the broader clinical impact of AI beyond just technical integration, addressing its potential for continuous learning and optimization in health care settings.

Response: Thanks for your review, this is well noted.

References

1. Anonymous. Peer review for "Checklist Approach to Developing and Implementing AI in Clinical Settings: Instrument Development Study". JMIRx Med 2025;6:e69869. [doi: 10.2196/69869]
2. Owoyemi A, Osuchukwu J, Salwei ME, Boyd A. Checklist approach to developing and implementing AI in clinical settings: instrument development study. JMIRx Med 2025;6:e65565. [doi: 10.2196/65565]
3. Zaki S. Peer review for "Checklist Approach to Developing and Implementing AI in Clinical Settings: Instrument Development Study". JMIRx Med 2025;6:e70058. [doi: 10.2196/70058]
4. Owoyemi A. Clinical AI sociotechnical framework (casof). Beadaut, Inc. URL: https://bit.ly/CASOF [accessed 2025-01-23]
5. Thompson K. Peer review for "Checklist Approach to Developing and Implementing AI in Clinical Settings: Instrument Development Study". JMIRx Med 2025;6:e69593. [doi: 10.2196/69593]
6. Saripalli S. Peer review for "Checklist Approach to Developing and Implementing AI in Clinical Settings: Instrument Development Study". JMIRx Med 2025;6:e69594. [doi: 10.2196/69594]

7. Nasa P, Jain R, Juneja D. Delphi methodology in healthcare research: how to decide its appropriateness. World J Methodol 2021 Jul 20;11(4):116-129. [doi: 10.5662/wjm.v11.i4.116] [Medline: 34322364]
8. Anonymous. Peer review for “Checklist Approach to Developing and Implementing AI in Clinical Settings: Instrument Development Study". JMIRx Med 2025;6:e69870. [doi: 10.2196/69870]
9. Anonymous. Peer review for “Checklist Approach to Developing and Implementing AI in Clinical Settings: Instrument Development Study". JMIRx Med 2025;6:e69595. [doi: 10.2196/69595]

Abbreviations

AI: artificial intelligence
CASoF: Clinical Artificial Intelligence Sociotechnical Framework

Edited by CN Hang, E Meinert, T Leung; submitted 02.12.24; this is a non–peer-reviewed article;accepted 02.12.24; published 20.02.25.

Please cite as:
Owoyemi A, Osuchukwu J, Salwei ME, Boyd A
Authors’ Response to Peer Reviews of “Checklist Approach to Developing and Implementing AI in Clinical Settings: Instrument Development Study”
JMIRx Med 2025;6:e69537
URL: https://xmed.jmir.org/2025/1/e69537
doi:10.2196/69537

Authors' Response to Peer Reviews of "Impact of a Point-of-Care Ultrasound Training Program on the Management of Patients With Acute Respiratory or Circulatory Failure by In-Training Emergency Department Residents (IMPULSE): Before-and-After Implementation Study"

Sandra Bieler[1], MD; Stephan von Düring[2], MD; Damien Tagan[3], MD; Olivier Grosgurin[4], MD; Thierry Fumeaux[5], MD, MBA

[1]Médecin cheffe, Service des Urgences, Hôpital de Sion, Sion, Switzerland

[2]Faculté de Médecine de l'Université de Genève, Hôpitaux Universitaires de Genève, Genève, Switzerland

[3]Service des Soins critiques, Hôpital Riviera Chablais, Rennaz, Switzerland

[4]Service de médecine interne générale et Service des Urgences, Hôpitaux Universitaires de Genève, Genève, Switzerland

[5]Hirslanden Geneva Clinics, Geneva, Switzerland

Corresponding Author:
Sandra Bieler, MD
Médecin cheffe, Service des Urgences, Hôpital de Sion, Sion, Switzerland

Related Articles:

Companion article: https://www.medrxiv.org/content/10.1101/2023.09.28.23295699v1

Companion article: https://med.jmirx.org/2025/1/e72144

Companion article: https://med.jmirx.org/2025/1/e53276

(JMIRx Med 2025;6:e72092) doi:10.2196/72092

KEYWORDS

point-of-care ultrasonography; training program; acute respiratory failure; acute circulatory failure; emergency department

This is the authors' response to peer-review reports for "Impact of a Point-of-Care Ultrasound Training Program on the Management of Patients With Acute Respiratory or Circulatory Failure by In-Training Emergency Department Residents (IMPULSE): Before-and-After Implementation Study."

Round 1 Review

Anonymous [1]

General Comments

This paper [2] researches an essential component of point-of-care ultrasonography. As this modality is rapidly evolving, evaluation of the impact on patient management and outcomes as well as cost-effectiveness is essential. Both subjects discussed in the paper result in a highly relevant manuscript. Even though the authors discuss relevant subjects, there are some problems with the manuscript.

Specific Comments

Major Comments

1. The title of the manuscript suggests that the authors researched the impact of ultrasound after implementation. However, as stated in the Methods section, ultrasound is already used by senior physicians. Thus, the impact of ultrasound after implementation is not researched but rather the impact of ultrasound used by residents. I suggest that the authors clarify that this is a feasibility and impact study on the implementation of point-of-care ultrasound (POCUS) used by residents in the emergency department (ED) in the title and Abstract.

Response: The title has been modified according to the reviewers' indications, to highlight the fact that the study's primary aim is to validate the implementation of a training curriculum for interns in training, and not to study the effect on patient outcome.

2. The authors state that patients were not included consecutively due to logistics in phase 2. This results in a high risk of bias in the included patients. Please include in the CONSORT (Consolidated Standards of Reporting Trials)

diagram the number of patients that were eligible and were excluded based on exclusion criteria or missed.

Response: As mentioned, the patients were not fully consecutively included due to organizational reasons: an incoming patient could only be considered for inclusion if the emergency department (ED) patient flow allowed, without delaying treatment or impacting on department operations. This is mentioned in the text. However, the number of patients who could have been included is not known (no traceability of screening).

3. It is unclear how many residents were performing the ultrasound examinations included in the analysis: the Methods section state that there was only 1 resident at the ED in both phases, while in the Results section, it states that there were 12 residents trained. Please clarify.

Response: Twelve doctors were trained, but only 1 resident at a time worked in the ED during each shift, and only he or she could therefore include patients during that shift, as specified in the text. We hope that the text will clarify this point.

4. The authors state that they chose a before-and-after implementation to evaluate the effect of POCUS to avoid contamination. However, before the implementation, POCUS was already used by senior physicians, which raises the question if POCUS was indeed not used in phase 1 of the trial.

5. Interestingly, in the Discussion section, the author discussed that the publication of Msolli et al did not find an improvement of diagnostic accuracy. It would be interesting to discuss why this is the case.

Response: As suggested by the reviewer, we have added a comment on the difference in the diagnostic accuracy of point-of-care ultrasound (POCUS) in our study and in the study by Msolli et al [3].

6. In the Discussion and Conclusion, it is suggested that the use of POCUS might lead to a decrease in hospital mortality. Since this is an observational study in which, just as the authors state, a diagnostic tool rather than a therapeutic intervention is researched, this is too rash to state. Please remove this from the Conclusion and Abstract.

Response: We have modified the Conclusion to relativize the effect of implementation on mortality, which is at best indirect, as mentioned by the reviewer.

Minor Comments

Overall

7. The authors provide results with IQR; however, no ranges are given. Please describe results as mean (SD) when data are normally distributed or median (25th percentile – 75th percentile) when data are not normally distributed.

Response: As all data are not normally distributed, we have chosen to keep the IQR (25th-75th), so as not to overload the text.

8. Formatting of the full manuscript needs some attention. For example, in the Abstract, not all sentences start with a capital letter. Also, it is common in the English language to write number in full up to 20.

9. Please follow the author guidelines of the journal for reporting values and the structure of the manuscript.

Response: Formatting has been adapted according to the transmitted comments.

Title Page

10. The authors state that a clinical trial registration was done. However, it seems that they refer to a registration by a medical ethical review board. Please provide a clinical trial registration or if not applicable, remove it from the title page.

Response: We have deleted the information on registration.

Introduction

11. In the first sentence, please state the full name of "emergency department" before using the abbreviation ED.

Methods

12. Figure 1 should be formatted. The most common formatting is according to the CONSORT flow diagram.

Response: We have formatted Figure 1 according to the instructions.

Results

13. Please do not discuss the results in the Results section.

Response: We have deleted all discussions of the results in the Results section.

Discussion

14. Please end the Discussion section with the strengths and limitations. The secondary findings should be above the Strengths and Limitations section.

Response: We have moved the secondary findings to before the discussion on the strengths and limitations.

Round 2 Review

Anonymous

I would like to compliment the authors of their extensive changes to the manuscript. I have some minor comments.

Response: We thank the editor and the reviewer for their careful reading of our manuscript and for their valuable comments. We have addressed all issues raised by them and modified the text accordingly. We have uploaded a change tracking version of the manuscript, with changes highlighted in yellow.

Before-and-after design: In such a study design, the only difference between the two phases should be the implemented intervention. In IMPULSE (Impact of a Point-of-Care Ultrasound Examination), the intervention was the implementation of immediate POCUS examination by junior in-training residents managing patients in the first line, after a short structured training program. This was performed only during the postimplementation phase, and never done before. POCUS could be performed in both phases by senior experienced physicians, but later in the management of the

patient, after the initial clinical evaluation (and after the POCUS during the postimplementation phase) of the junior resident. We therefore continue to affirm that this is indeed a before-and-after study design, with a clear implementation of a changing practice. We have clarified this in all sections of the text.

We have, as suggested, included information on the residents' characteristics, as this valuable information is important for the interpretation of the study results. A new section has been added in the Methods and in the Results parts of the text.

We have put the 25th - 75th IQR range everywhere in the text and tables, as suggested.

We have removed the figure legends from the uploaded figures.

As mentioned, a change-tracking version has been uploaded as a supplementary file, with changes highlighted in yellow.

All ethics information has been grouped in a specific section in the Methods part of the text.

We have followed the guidelines on reporting results.

Minor Comments

1. I would suggest changing the sentence "However, there is still no strong evidence that the diagnostic accuracy of POCUS translates into a clinically relevant difference in patient outcomes" in the Introduction, because you also do not provide strong evidence (I do not know if we ever could provide strong evidence). I would suggest that you focus it more on the fact that the impact of using POCUS in the management of patients in the ED is still relatively unknown.

Response: We have adapted the sentence on the evidence of the clinical impact of POCUS in the Introduction, as suggested by the reviewer.

2. I would suggest to start the Discussion section with a short summary of the key findings.

Response: We have started the Discussion section with a short summary of key findings.

References

1. Anonymous. Peer review of "Impact of a Point-of-Care Ultrasound Training Program on the Management of Patients With Acute Respiratory or Circulatory Failure by In-Training Emergency Department Residents (IMPULSE): Before-and-After Implementation Study". JMIRx Med 2025;6:e72144. [doi: 10.2196/72144]
2. Bieler S, Tagan D, Grosgurin O, Fumeaux T. Impact of a point-of-care ultrasound training program on the management of patients with acute respiratory or circulatory failure by in-training emergency department residents (IMPULSE): before-and-after implementation study. JMIRx Med 2025;6:e53276. [doi: 10.2196/53276]
3. Msolli MA, Sekma A, Marzouk MB, et al. Bedside lung ultrasonography by emergency department residents as an aid for identifying heart failure in patients with acute dyspnea after a 2-h training course. Ultrasound J 2021 Feb 9;13(1):5. [doi: 10.1186/s13089-021-00207-9] [Medline: 33559777]

Abbreviations

ED: emergency department
IMPULSE: Impact of a Point-of-Care Ultrasound Examination
POCUS: point-of-care ultrasound

Edited by E Meinert, A Schwartz; submitted 03.02.25; this is a non–peer-reviewed article;accepted 03.02.25; published 03.03.25.

Please cite as:
Bieler S, von Düring S, Tagan D, Grosgurin O, Fumeaux T
Authors' Response to Peer Reviews of "Impact of a Point-of-Care Ultrasound Training Program on the Management of Patients With Acute Respiratory or Circulatory Failure by In-Training Emergency Department Residents (IMPULSE): Before-and-After Implementation Study"
JMIRx Med 2025;6:e72092
URL: https://xmed.jmir.org/2025/1/e72092
doi:10.2196/72092

Authors' Response to Peer Reviews of "Predicting Escalation of Care for Childhood Pneumonia Using Machine Learning: Retrospective Analysis and Model Development"

Oguzhan Serin[1], MD; Izzet Turkalp Akbasli[1], MD; Sena Bocutcu Cetin[1], MD; Busra Koseoglu[1], MD; Ahmet Fatih Deveci[2], MSc; Muhsin Zahid Ugur[2], PhD; Yasemin Ozsurekci[3], MD

[1]Department of Pediatrics, Hacettepe University Medical School, Gevher Nesibe Avenue, Altindag, Ankara, Turkey

[2]Department of Health Information Systems, University of Health Sciences, Istanbul, Turkey

[3]Department of Pediatric Infectious Diseases, Hacettepe University Medical School, Ankara, Turkey

Corresponding Author:
Izzet Turkalp Akbasli, MD
Department of Pediatrics, Hacettepe University Medical School, Gevher Nesibe Avenue, Altindag, Ankara, Turkey

Related Articles:

Companion article: https://www.medrxiv.org/content/10.1101/2024.02.22.24303209v1

Companion article: https://med.jmirx.org/2025/1/e71100

Companion article: https://med.jmirx.org/2025/1/e71369

Companion article: https://med.jmirx.org/2025/1/e57719

(JMIRx Med 2025;6:e71098) doi:10.2196/71098

KEYWORDS

childhood pneumonia; community-acquired pneumonia; machine learning; clinical decision support system; prognostic care decision

This is the authors' response to peer-review reports for "Predicting Escalation of Care for Childhood Pneumonia Using Machine Learning: Retrospective Analysis and Model Development."

Round 1 Review

Anonymous [1]

General Comments

This paper [2] developed a machine learning approach that could predict community-acquired pneumonia prognosis, which is scaled into two-levels, severe or nonsevere, and identify important clinical indices, such as hypoxia, respiratory distress, age, z score of weight for age, and antibiotic usage before admission. The machine learning–based clinical decision support system tool for childhood pneumonia could provide prognostic support for case management.

Response: Thank you for your positive summary of our work. We appreciate your recognition of the machine learning tool's potential in supporting childhood pneumonia prognosis and case management.

Specific Comments

Major Comment

1. To enhance the manuscript's grounding in current research and to provide a comprehensive context for the study, the authors are recommended to incorporate an evaluation of related literature in the Introduction and Discussion sections. This could include, but not be limited to, the following studies:

- *Liu YC, Cheng HY, Chang TH, et al. Evaluation of the need for intensive care in children with pneumonia: machine learning approach. JMIR Med Inform. Jan 27, 2022;10(1):e28934. [doi: 10.2196/28934] [Medline: 35084358]*
- *Smith JC, Spann A, McCoy AB, et al. Natural language processing and machine learning to enable clinical decision support for treatment of pediatric pneumonia. AMIA Annu Symp Proc. Jan 25, 2020;2020:1130-1139. [Medline: 33936489]*
- *Kanwal K, Khalid SG, Asif M, Zafar F, Qurashi AG. Diagnosis of community-acquired pneumonia in children using photoplethysmography and machine learning-based classifier. Biomed Signal Process Control. Jan 2024;87:105367. [doi: 10.1016/j.bspc.2023.105367]*
- *Chang TH, Liu YC, Lin SR, et al. Clinical characteristics of hospitalized children with community-acquired*

pneumonia and respiratory infections: Using machine learning approaches to support pathogen prediction at admission. J Microbiol Immunol Infect. Aug 2023;56(4):772-781. [doi: 10.1016/j.jmii.2023.04.011] [Medline: 37246060]

The readers could have a more comprehensive understanding if the authors could include a concise evaluation of the prior literature in the current manuscript.

Response: Thank you for those invaluable articles. We have revised the Introduction and Discussion sections to include a concise evaluation of the recommended studies, along with other relevant literature, in order to enhance the readers' understanding and to enhance alignment with the current research landscape in this niche.

2. Considering the high stakes involved in pediatric care, particularly in intensive settings, it is critical to exam the false negative cases from the confusion matrices. Analyzing these cases for any common feature characteristics could provide insights into potential improvements in the predictive algorithm. This analysis should be clearly presented and discussed in the manuscript, emphasizing its importance in clinical decision-making.

Response: Thank you for this important suggestion. We have carefully reviewed the false negative cases and conducted an analysis to identify any common characteristics. The analysis of false negatives of the best model "Blending-2" only revealed two false negatives, underweighting clinical features comorbidities while over-relying on the absence of hypoxia. As it only included two cases, the false negatives analysis has not been included in the Results section.

3. The manuscript would benefit from a more detailed description of the cohort used in the study. Information on age, gender, and other clinical indices across the two groups (severe and nonsevere) would enable a better understanding of the study population. Additionally, providing the number of cases in each group would clarify the scope and scale of the study findings.

Response: We have added a Study Population section in the Methods, providing details on the study group and the candidate variables collected. Additionally, a Study Population Characteristics section has been included in the Results, where key variables (eg, age, respiratory distress, and leukocyte count) are compared between the nonsevere and severe level of care groups (Table 2). These updates clarify the cohort's characteristics and address your concern regarding study population details.

4. A detailed description of the data collection process is crucial for assessing the study's applicability in real-world clinical settings. The manuscript should explicitly state the following:

- *How and when clinical data, including features such as hypoxia and respiratory distress, were collected (eg, at the time of admission? or within 24 hours of admission?);*
- *The time frame considered for "antibiotic usage before admission" as relevant to the prediction model: This information is essential for replicability and for future applications of the findings in clinical workflows.*

Response: We have provided a detailed description of the variables in the revised Table 1 to enhance transparency, ensuring a better understanding of how data were collected and used for the prediction model. All clinical features were encoded by pediatricians using the unstructured initial medical records at admission. For clarity and the comprehension of readers, the phrase "...candidate features from unstructured admission notes" was added to the second paragraph under the subheading of Case Definition and Patient Selection in the Methods section. Additionally, The term "recent antibiotic usage" has been clarified to indicate oral antibiotic use prescribed before admission, specifically within the 14 days preceding hospitalization. We believe these additions provide the necessary clarity and improve the replicability of the study in real-world clinical workflows.

Reviewer E [3]

General Comments

The authors have examined the medical records for 437 patients with pneumonia and created a machine learning–based classifier to determine which patients required transfer to a tertiary care center. This subject is interesting, as the predictive power of these novel statistical techniques is high and could improve the clinical care of these patients. The authors have done thorough work describing the statistical methods used in the preprocessing of the data and model development. My primary concerns in the manuscript are the lack of clinical application description, the lack of description of the time frame of the included data elements, and the lack of description regarding the patient population and outcome of interest. The following are my point-by-point comments.

Response: Thank you for your thoughtful and detailed review of our manuscript. We appreciate your recognition of the statistical methods we used for preprocessing and model development. We acknowledge the need for improving our work in the fields that Reviewer E stated. Therefore, we have addressed each of these points as follows:

- The updated Table 1 (candidate features) provides an in-depth description of the clinical and laboratory features on how and when data collection was made (time frame), along with their clinical relevance in predicting the outcome of level of care severity. These variables were chosen based on their clinical value and ease of collection in primary care settings, allowing the model to be functional in low-resource environments.
- A new Table 2 (former Table 2 became Table 3) presents a statistical comparison between the severe and nonsevere level of care groups, focusing on the differences in demographics, clinical presentation, and laboratory values. This further highlights the factors that contribute to the outcome of interest—whether a patient requires tertiary care. The revised tables should provide a more comprehensive understanding of how the model was developed and how it applies to real-world clinical populations.

- A new subsection titled Study Population Characteristics was added under Results, where key variables were compared between groups, along with presenting the characteristics of the study population.

Specific Comments

Major Comments

Abstract

The authors use the term "case management" in the Abstract and several times in the manuscript. In this context, the authors' meaning is the decision for the escalation of care or patient transfer. However, in US-based hospital systems, case management has a different meaning, which includes largely transition to rehabilitation or nursing facilities, acquisition of home oxygen therapy, etc. I would recommend altering this term for comprehension to something like "escalation of care" or "patient triage."

Response: We acknowledge that the term "case management" may have different interpretations depending on the health care system. To avoid confusion, we will revise this term throughout the manuscript (including the main title) to either "prognostic care decision," "diagnosis and treatment," or "pneumonia management," which are more in alignment with our study's goal and contemporary research. Additionally, the Abstract has been substantially revised to align with the updated version of the manuscript.

The primary outcome of interest should be included in the Abstract.

Response: We have included a clear statement in the Abstract that the primary outcome of interest is the level of care severity, specifically focusing on the need for pediatric intensive care unit admission or advanced respiratory support.

As detailed in the Methods section, it is crucial to describe the time frame for the included variables, to know when the algorithm could be used in clinical practice.

Response: We specified the time frame for the data collection in the Abstract, in alignment with the changes made in texts and tables in the Methods section, ensuring that readers understand when the algorithm could be used in clinical practice. This will clarify the applicability of the model based on the retrospective nature of the data.

Introduction

As the goal of the algorithm in the study is to predict which patients will need transfer to tertiary care for increasing respiratory support, more of the Introduction should focus on the management of in-hospital pediatric pneumonia, challenges, and reasons for the escalation of care.

I would recommend altering the sentence that describes pneumonia as easily preventable and treatable. Several of the most complicated cases in the intensive care unit are admitted with pneumonia.

Response: Thank you for your valuable suggestions regarding the focus of the Introduction. We have revised the section to better emphasize the management of in-hospital pediatric pneumonia, including the challenges faced in recognizing and managing disease severity, as well as the reasons for escalating care. Furthermore, we have altered the sentence describing pneumonia as "easily preventable and treatable" to acknowledge the complexity of cases, particularly in intensive care settings. The revised Introduction includes the following:

1. Challenges and reasons for the escalation of care: To address this suggestion, we have expanded on the reasons for the escalation of care, providing the literature standpoint for the reasons of selecting candidate features.
2. Clarification of pneumonia's preventability and treatability: We have revised the sentence that previously described pneumonia as "easily preventable and treatable" to better reflect the complexity of the disease.
3. More focus on the management of in-hospital pediatric pneumonia: With all respect to this comment, we kindly disagree to have more focus on in-hospital pneumonia care, as it would shift the main objective of this study, which is providing prognostic care tools for primary care settings.

Methods

While great care is taken to describe the approach to data preprocessing, feature selection, and model development, I would recommend following the TRIPOD (Transparent Reporting of a Multivariable Prediction Model for individual Prognosis or Diagnosis) guidelines [4], which are validated reporting recommendations for predictive models.

Response: Thank you for the insightful suggestion. We have reviewed the TRIPOD checklist and ensured that our manuscript adheres to these guidelines for transparent reporting of predictive models. We have uploaded the filled checklist under the section of "Upload Additional Material (for editors/reviewers' eyes only)."

Please provide more details regarding the hospital systems involved in this study. Are they large, academic centers or small, rural centers?

Response: Thank you for your insightful comment. In response, we have clarified the institution in the Methods section to provide better context on the hospital system involved.

For study inclusion, I am not familiar with the Integrated Management of Childhood Illness guidelines. Are these structured diagnostic codes captured in the electronic health record? Is it a computational phenotype?

Response: Thank you for raising this important point. The Integrated Management of Childhood Illness guidelines are World Health Organization recommended, providing a clinical framework for diagnosing and managing pneumonia, but they are not structured diagnostic codes in the electronic health record. Physicians manually encoded clinical features from unstructured admission notes for phenotyping, rather than using a computational phenotype. This clarification has been added to the Methods section.

Please specify what is meant by "neonatal age."

Response: We appreciate your suggestion for greater clarity. We have now specified that "neonatal age" refers to infants

younger than 28 days of life. This has been updated in the Methods section for precision.

Many of the variables included in the model are colinear. For example, age and weight are highly dependent on one another, and including both in the model can be detrimental. The feature selection methods may be able to discern this, but maybe not. I would recommend using only age and z score in the model.

Response: We appreciate your insightful comments and suggestions. It appears that including both "weight" and the "weight-for-age *z* score" derived from national reference values based on age may have caused some confusion. We have clarified this issue to ensure a more coherent presentation of the candidate features. As we only included the weight-for-age *z* score (and not weight in kilogram) in our first model, no further adjustment is required in this regard. We have retained "age" as a feature because respiratory infections and disease characteristics can vary significantly across age groups. Additionally, we kept "weight-for-age *z* score" as a separate variable, as it reflects the child's relative position among peers in the nation and serves as an indirect indicator of nutritional status.

The time frames are not stated for the variables. For example, does "hypoxia" mean hypoxia at any time during the hospitalization? On hospital admission? In the first 12 hours? This information is vital to determine the usability of the entire model. If the model uses variables available during the entire hospitalization, the predictive ability will be high, but the usability will be low. A model that can predict right when a patient is transferred to a tertiary care center that the patient will be transferred is useless. However, a model that can predict on admission, or in the first 6 - 12 hours, that a patient will require transfer is incredibly helpful. Without knowing the time frame for these variables, we cannot assess how the model could be applied in clinical practice.

Response: We thank both reviewers for raising this important point. We agree that specifying the time frames for the variables is crucial for understanding the model's applicability in clinical settings. In response, we have clarified the data collection process in the revised manuscript. All clinical features, including hypoxia and respiratory distress, are now detailed in the updated Table 1 and additional text in the Methods section under Case Definition and Patient Selection, with more emphasis on the relevant time frames of the features.

Please provide clarity regarding the study outcomes. The primary outcome is described as whether the patient was referred to a tertiary care center or not. The next sentence describes "poor prognosis" as pediatric intensive care unit admission or oxygen/ventilation support. How is this outcome used? Is this a secondary outcome? Is this describing the reason for transfer? Please clarify.

Response: Thank you for highlighting this point. We acknowledge the need to clarify the study outcomes. The primary outcome is whether the patient requires transfer to a tertiary care unit. The term "poor prognosis" refers to the reason for transfer, specifically whether the patient required pediatric intensive care unit admission or oxygen/ventilation support. This is not a separate secondary outcome, but rather the criteria used to define the primary outcome of requiring tertiary care. We have revised the manuscript to clarify that the primary outcome is the "Level of Care Severity," along with text in the Methods section to make this distinction clear.

As stated in the TRIPOD guidelines, you should present the amount of missingness in your data. It appears you used imputation methods for missing data. It is helpful to describe the amount of missing data that was imputed and the method for imputation.

Response: Thank you for your valuable comment. In accordance with the TRIPOD guidelines, we agree that reporting the amount of missing data is important for transparency. We should have mentioned our imputation method while providing details about relevant features in the first submission. We have now included a detailed description of the missing data in our revised manuscript, specifying both the percentage of missing values for each variable and the total amount of missing data. To handle missing data, we used the light gradient boosting machine algorithm as an imputation method, treating missing values as a dependent variable and predicting them based on other features to avoid bias. Individual feature weights were applied accordingly. The following features had missing values: C-reactive protein (n=34, 8.2%), albumin (n=10, 2.4%), sodium (n=8, 1.9%), aspartate aminotransferase (n=16, 3.9%), and alanine aminotransferase (n=16, 3.9%). This information has been added to the revised manuscript for clarity.

Results

There is a glaring lack of information regarding your study population. Please provide a table describing patient characteristics including demographics and the variables you used in the algorithm. Also, please provide a comparison between the patients who were transferred to a tertiary care center and those who were not.

Response: Thank you for your observation. In response, we have added a detailed description of the study population in the revised manuscript. Specifically, we have included a new subsection titled Study Population Characteristics, along with a new Table 2, which presents a comparison of the demographic and clinical characteristics between the severe and nonsevere level of care groups. We have also used appropriate statistical tests to compare the characteristics of patients requiring transfer to a tertiary care unit (severe care group) versus those who did not (nonsevere group). These additions enhance the clarity of our population description and provide a comprehensive comparison of the key variables used in our algorithm.

In imbalanced datasets, it can be more useful to measure model performance using the area under the precision-recall curve rather than the standard area under the receiver operator characteristic curve. I would recommend adding this metric.

Response: Thank you for your insightful suggestion. We agree that in the case of imbalanced datasets, the area under the precision-recall curve (PRC) can provide a more informative measure of model performance than the standard area under the receiver operating characteristic curve. In response, we have now added the PRC of all models in the performance table. We

also included a PRC plot for the blending model labeled as "Blending-2," which incorporates the top-5 highest-ranked clinical features using the optimized CatBoost, light gradient boosting machine, and extreme gradient boosting models. The new PRC plot, along with the text explaining it in the Results section, have been added to the supplementary materials to provide a more comprehensive evaluation of the model's performance on imbalanced data.

Discussion

The Discussion, overall, focuses much more on the technical details of the data curation and model development than it does on the clinical application of the model. Much of the technical details presented are also clearly explained in the Methods section and then repeated in the Discussion. I would recommend substantial revision to the Discussion section to remove redundant information that is already contained in the Methods section, as well as the addition of how this model could be applied in a clinical setting to improve the care of patients with pneumonia.

Response: We thank the reviewer for this valuable feedback. In response, we have thoroughly revised the Discussion section to reduce redundancy and place a greater focus on the clinical applications of the model, along with contemporary study inclusion. Specifically, we removed technical details that were previously repeated from the Methods section, such as the handling of imbalanced data with Synthetic Minority Oversampling Technique–Tomek, feature selection using Shapley additive explanations and recursive feature elimination with cross-validation, and detailed performance metrics for each algorithm.

In place of these technical details, we have expanded the Discussion to focus more on how the model can be used in a clinical setting to improve pneumonia care. We now highlight how the model can assist primary care physicians, especially those working in resource-limited environments, in identifying high-risk pneumonia cases that may require referral to tertiary care. We also put emphasis on predictive features (such as hypoxia, respiratory distress, age, weight *z* score, and complaint period) that are easy to assess in primary care, making the model highly practical for use in real-world clinical settings. Furthermore, we discuss the potential for the model to improve patient outcomes by facilitating timely care decisions, particularly in settings where advanced diagnostic tools may not be available.

The Discussion contains no information regarding the limitations of the study. Please describe in detail the prominent limitations of the study. These should include the use of retrospective data, including only two centers, imbalanced data, challenges with clinical implementation of the model, etc.

Response: Thank you for highlighting the need to discuss the limitations of the study in more detail. In response, we have expanded the Discussion section to include a more comprehensive account of the study's limitations. Specifically, we now address the reliance on data from a single tertiary hospital, the potential selection bias toward severe cases, the limited sample size, and the retrospective nature of the data.

The Discussion, and other areas of the manuscript, mention disease prevention several times. The goal of this study has nothing to do with the prevention of pneumonia, only the treatment of pneumonia and the prevention of associated morbidity and mortality. Please revise.

Response: Thank you for pointing out the unnecessary mentions of disease prevention in the manuscript. We agree that the primary focus of the study is on the treatment of pneumonia and the prevention of associated morbidity and mortality, not the prevention of the disease itself. We have revised the entire manuscript to eliminate any mention of disease prevention where it is not relevant and have ensured that the discussion stays focused on treatment and prognosis.

Conclusion

As it stands, the Conclusion is fairly long and does not focus only on the primary findings of the study. I would recommend trimming it to 2 - 3 sentences that focus only on the primary findings of the study, such as the feasibility of developing this type of predictive model and the potential applications of the model to clinical practice.

Response: Thank you for your feedback regarding the length and focus of the Conclusion. We agree that the Conclusion could be more concise and focused on the primary findings. Based on your suggestion, we have significantly shortened the Conclusion to focus solely on the primary findings of the study, namely, the feasibility of developing a predictive model for childhood pneumonia prognosis and its potential clinical applications. The revised Conclusion now highlights the key outcomes concisely.

Minor Comments

Methods

The authors describe that ensemble methods "significantly enhance the accuracy of classifications." Please provide a reference for this statement.

Response: We agree that providing a reference would strengthen this statement. We have now included a reference supporting our statement. Specifically, "Mahajan P, Uddin S, Hajati F, Moni MA. Ensemble learning for disease prediction: a review. Healthcare (Basel). Jun 20, 2023;11(12):1808. [doi: 10.3390/healthcare11121808] [Medline: 37372925]"

Results

Please provide numbers for those who met your primary outcome of interest (transfer to a tertiary care center).

Response: Thank you for your suggestion to provide specific numbers related to the primary outcome of interest. We have now revised the Results section to include study population characteristics along with a comparison between the severe (transferred to a tertiary care unit) and nonsevere level of care groups. The revised Results section also holds emphasis on the primary outcome of interest as follows "...Of the 437 patients analyzed, 304 patients (69.6%) met the primary outcome of being transferred required escalation of care."

Please provide a description of the time frame for patient transfer, for those who were transferred.

Response: In alignment with previous comments on the inclusion of time frames to relevant data elements, we have provided a detailed description in the updated Table 1 for candidate variables. However, our dataset does not include the timing of transfers to tertiary care units. This is recognized as a limitation of the study, and the Limitation section has been extended in this regard.

Discussion

It would be interesting to hear more regarding the use of this model in resource-limited settings and the benefits it could provide.

Response: Thank you for your valuable comments, which have already enhanced our work beyond our initial vision. We share your excitement about the future potential of this work and its possible applications.

Round 2 Review

Anonymous

I thank the authors for revising the manuscript.

Reviewer E

General Comments

The authors have conducted a single-center, retrospective study evaluating the derivation and performance of a machine learning model to predict the need for transfer to a higher level of care for childhood pneumonia. The authors were provided with a substantial amount of feedback on the original submission, and although the authors' response is detailed and comments on how all concerns were adequately addressed, the resulting manuscript is lacking in many if not most of the requested changes. The revised manuscript remains confusing to the reader and bereft of some essential elements of standard study reporting, including a basic description of the patient population and details regarding the timing of variable collection and use in the model. Due to this lack of response to the initial reviewer feedback, I am recommending rejection of this manuscript. The following are my point-by-point critiques, many of which are similar to those in my original review.

Response: We believe that these comments may stem from a review of the earlier version of our manuscript rather than the revised submission. Each specific comment raised by the reviewer was addressed in the revised manuscript, where we carefully incorporated the requested changes and clarifications. We kindly request a review of the latest version in the JMIRx system, as it reflects these substantial updates in response to the initial feedback. As the reviewer provided some additional recommendations, we made the required changes to those in our most recent manuscript. We believe there may have been a misunderstanding or an oversight, leading to the reviewer evaluating an earlier version of our manuscript. We genuinely appreciate the time and effort the reviewer has invested in helping us improve our manuscript.

Specific Comments

Abstract

First sentence: Please revise it to "Pneumonia is the leading cause of preventable mortality for children under five years of age."

Response: We have revised the first sentence of the Background section of the Abstract.

Background: The terms "case management" and "disease prevention" are still used in the Abstract. In my initial review, I recommended revising these terms to improve study clarity, and although the authors stated in their response that they replaced these terms, they remain in the Abstract. As it stands, it is not immediately clear to the reader that the purpose of the study was to provide a tool to assist bedside clinicians to determine which patients are likely to require transfer of care to a higher-level facility for pediatric pneumonia.

Response: Thank you for highlighting the importance of precise terminology in conveying the study's purpose. We have already revised the entire document to address the reviewer's initial comment/concern. We have now double-checked the revised manuscript and there is no mention of "case management" in the revised manuscript, as well as "disease prevention," that could be misunderstood by readers.

Methods: As it stands, it is confusing to the readers what was actually done in the study. It should be very apparent that the authors used a specific list of variables (please provide each in the Abstract) to predict the need for transfer to a larger institution using a specific type of machine learning model (ensemble). In the current version, this is difficult to discern.

Response: We thank your attention to the need for clarity in the Abstract. We have already addressed this concern by stating "Pediatricians encoded key clinical features from unstructured medical records based on IMCI guidelines." This line conveys that essential variables were derived from standardized guidelines without detailing each variable. Listing all variables in the Abstract would reduce clarity when considering the Abstract word limitations of this journal, especially since these variables are fully detailed in the Methods and Results sections. We believe this approach aligns with best practices for Abstract conciseness and provides sufficient information for the reader.

Results: I would be completely clear regarding the outcome your model is predicting. After reading the paper, it is understood that "pneumonia prognosis" and "severity" actually mean required transfer to a higher level of care, but it is unclear in the Abstract. I would explicitly state "predicted transfer to a higher level of care with 77% - 88% accuracy."

Response: Thank you for this valuable suggestion to improve clarity. In response, we have revised the Results section of the Abstract to explicitly state that the model predicts the need for transfer to a higher level of care, specifying the accuracy range as suggested. The revised phrasing is now "The optimized models predicted the need for transfer to a higher level of care with an accuracy of 77% - 88%..." This adjustment enhances clarity and directly conveys the model's intended outcome for readers.

Introduction

Second paragraph, fifth sentence: I would recommend revising it to "However, this preventable health problem continues to be a substantial cause of mortality, especially in underdeveloped countries and regions, due to the lack of equipment and trained human resources." There is no way to quantify it as "the most important cause of mortality."

Response: There is no mention of "the most important cause of mortality" in the revised manuscript. However, we noticed that it was in the first submission. We are deeply concerned that the reviewer's second round of comments did not provide feedback on the revised manuscript.

The term "case management" continues to be used in the Introduction, which decreases clarity for the reader.

Response: Again, these concerns have already been addressed in the revised manuscript. There is no mention of "case management." We kindly request the reviewer to read the revised version rather than the first submission that has been substantially changed after the reviewer's initial comments.

As recommended previously, I would be very specific in the Introduction that you are trying to create a tool to help bedside clinicians (typically non–intensive care physicians) decide when to transfer a patient with pneumonia to a higher level of care to prevent morbidity and mortality. As it stands, this is unclear.

Response: Thank you for this recommendation. This point was already addressed in the revised manuscript, where we clarified the study's goal in the Introduction. Please also refer to the Introduction section in the last paragraph, stating "We aimed to develop machine learning-based clinical decision support system tool for childhood pneumonia that can be used by physicians, particularly working in LMICs." However, we believe including the adjective "non-intensive care" to define these physicians in detail would improve the manuscript.

Methods

In my initial review, I asked the authors to clarify what is meant by neonatal age. In their response, they said they had revised the Methods to state specifically 28 days or fewer. However, in the first paragraph of the Methods, it continues to state "neonatal age." Please revise.

Response: Thank you for raising this point again. We did agree on this issue and corrected it in the revised manuscript as follows: "Patients younger than 28 days of age (neonatal age), older than 18 years, and those who had been hospitalized within the last 14 days were excluded." Preserving the neonatal age in this sentence is essential to emphasize that we are excluding newborn pneumonia, which requires way different clinical management and decisions.

For clarity, I would recommend restating your primary outcome to simply "required tertiary care referral." Having the outcome as severe versus nonsevere, which is defined as requiring tertiary care referral or not, adds an extra step to the thought process and can be confusing.

Response: We appreciate the recommendation to clarify the primary outcome. In the revised manuscript, we have already redefined the primary outcome to "Level of Care Severity," scaled as severe or nonsevere, and defined it as the need for referral to a tertiary care unit for intensive care or respiratory support. This phrasing preserves the conceptual framework of care severity levels while directly specifying that the outcome reflects the requirement for tertiary care referral. We believe this approach balances clarity with the study's structured outcome definitions. Additionally, this terminology is consistently used in the entire manuscript, including the Methods section, where we explicitly defined it in Table 1.

One of my largest concerns in the initial manuscript was the timing of the variables. This is crucial when determining how useful the model could be. If the elements in Table 1 are measured on admission, or in the first 6 - 12 hours of admission, the model could be very useful for patient care. If the elements were measured at any point during the hospitalization, it becomes much less useful. My worry is that the model was developed based on the elements' presence at any point, meaning if the child had fever, cough, respiratory distress, and hypoxia at hour 48, then at hour 49 the model was able to predict the patient would need transfer, and the patient was transferred at hour 50—this is not helpful to clinicians. On the other hand, if the model predicts at hour 12 that a patient needs transfer, and then at hour 50 they transfer, that is potentially very helpful to clinicians. Without these details, I cannot recommend the publication of the manuscript.

Response: Thank you for emphasizing the importance of timing in assessing the model's clinical utility again. We have already clarified this point in the revised manuscript by specifying that all variables in Table 1 were recorded at the time of admission. As stated in Table 1, these variables were extracted from initial examination documents, not from any time from the hospitalization period, reflecting the presence/measurement of variables at admission. We believe that timings are adequately mentioned by the "at admission" or "at initial examination" phrases in Table 1. Only the primary outcome "Level of Care Severity" was extracted from medical records other than the initial time point, as it is necessary to encode whether or not a patient had advanced support during their hospital stay.

It appears that the model was developed using the data from all 437 patients, and the results are presented following k-fold cross validation. It is standard practice to derive the model on a subset of the data (typically 70% - 80%) and then to test it on the remainder of the dataset to prevent overfitting and inflation of performance metrics. It does not appear that this was done. Despite having a small sample size, I believe this approach would lead to a more robust and generalizable model.

Response: Thank you for highlighting this point regarding model validation. In the revised manuscript, we confirmed that a k-fold cross-validation approach was used on the entire dataset to address the limited sample size. To mitigate concerns of overfitting and enhance model generalizability, we initially split the data, setting aside 5% as a test set to prevent data leakage. The remaining data were then used in an 85%:15% split for training and validation. This approach was chosen to maximize the utility of our sample while ensuring a robust evaluation of model performance. Please refer to the subsections named

Handling With the Imbalanced Dataset and Algorithms, where we have already addressed the reviewer's concern, in the revised manuscript from the round 1 review.

Results

The first paragraph contains many "nuts and bolts" details of model development, and these would be better positioned in the Methods section.

Response: Again, we are deeply concerned that the reviewer may not be reading the revised manuscript from the round 1 review. These concerns have already been addressed. In the revised manuscript, the Results section begins with subsection named Study Population Characteristics.

Both reviewers on the initial submission requested additional details describing the study population, and although the authors responded that they added these details, there are still none provided. It is essential to the understanding of the study results to know the characteristics of the patient population, and it should be a standard requirement for all clinical studies.

Response: We have already agreed on this issue and carefully included a substantial revision with a Study Population Characteristics subsection and a detailed Table 2, reflecting the study population adequately. Please refer to these sections, and we are prepared to address any further concerns regarding the presentation of the study population if needed.

The Shapley additive explanations value results presented in Figure 2 are valuable, but more details describing each measured factor are required. I recommend a table with each factor as rows and two columns comparing the population that did not require transfer to a tertiary care center to the population that did.

Response: Again, this concern has already been addressed by Table 2, with a basic statistical comparison between two groups including test statistics with the significance level.

An additional figure showing an area under the precision-recall curve for each model would also be interesting to the readers.

Response: On the round 1 revision, we have already included a new figure in Multimedia Appendix 2, showing the PRC. This may have been spared from the reviewer's eye.

Discussion

The Discussion spends a decent amount of space discussing the COVID-19 pandemic. While this does have some bearing on the management of childhood pneumonia, I believe the space would be better spent discussing the actual implementation of this type of algorithm. How would a primary care clinician actually use this model in practice? How would it improve upon current clinical practice? Would it be easy or difficult to incorporate into routine workflows? This would be more interesting to the readers.

Response: The revised manuscript has substantially been changed, reducing the amount of emphasis on the pandemic and carefully answering those questions that have been raised by the reviewer in the first round.

I recommend adding what the next steps of this line of research would be. How would you seek to improve the model's performance? More patient data? Additional variables?

Response: We have provided recommendations along with our limitations. Please refer to our Limitation paragraph—specifically, just before the Conclusion paragraph.

In the original submission, I recommended the authors provide a limitations section and also provided some examples. Although the authors response says they added this, there are still no limitations provided. Please provide this essential element to the Discussion.

Response: This new comment provides evidence that the reviewer was not reading the revised manuscript from the first round, because we have one relatively long paragraph dedicated to the limitations of this study. The Limitation paragraph starts with "One significant limitation of this study…" We have double-checked the JMIRx submission system, and we confidently confirm that we have uploaded the revised manuscript correctly.

Conclusion

I recommend commenting on what the next steps of this line of research would be in more specific terms.

Response: We believe that our Conclusion reflects the primary findings of the study along with its clinical importance and applicability.

Round 3 Review

Reviewer E

General Comments

The authors have conducted a single-center, retrospective study evaluating the derivation and performance of a machine learning model to predict the need for transfer to a higher level of care for childhood pneumonia. The authors were provided with a substantial amount of feedback on the original submission and have been responsive to feedback, which has resulted in a much improved manuscript. There remain several typographical and grammatical errors, which I would advise an English-grammar expert to review prior to publication, but from a scientific standpoint, I believe the manuscript is appropriate for publication.

Response: We sincerely appreciate the reviewer's recognition of the improvements made to the manuscript and their support for its scientific merit. We have carefully reviewed the manuscript for typographical and grammatical errors to ensure the highest standard of clarity and professionalism prior to publication. Thank you again for your valuable feedback that improved the quality of our work.

Specific Comments

Major Comments

1. *Details regarding the patient population have been provided in detail.*
2. *The study objectives have been clarified for readers.*
3. *The study methods are now much more reproducible.*

Response: These aspects were prioritized during the revision process, guided by the reviewers' constructive feedback, which significantly enhanced our work. Their insightful comments not only improved this manuscript but also provided valuable lessons for our future works.

References

1. Anonymous. Peer review of "Predicting Escalation of Care for Childhood Pneumonia Using Machine Learning: Retrospective Analysis and Model Development". JMIRx Med 2025;6:e71369. [doi: 10.2196/71369]
2. Serin O, Akbasli IT, Cetin SB, et al. Predicting escalation of care for childhood pneumonia using machine learning: retrospective analysis and model development. JMIRx Med 2025;6:e57719. [doi: 10.2196/57719]
3. Rogerson C. Peer review of "Predicting Escalation of Care for Childhood Pneumonia Using Machine Learning: Retrospective Analysis and Model Development". JMIRx Med 2025;6:e71100. [doi: 10.2196/71100]
4. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent Reporting of a Multivariable Prediction Model for individual Prognosis or Diagnosis (TRIPOD): the TRIPOD statement. BMJ 2015 Jan 7;350:g7594. [doi: 10.1136/bmj.g7594] [Medline: 25569120]

Abbreviations

PRC: precision-recall curve
TRIPOD: Transparent Reporting of a Multivariable Prediction Model for individual Prognosis or Diagnosis

Edited by E Meinert, S Amal; submitted 09.01.25; this is a non–peer-reviewed article;accepted 09.01.25; published 04.03.25.

Please cite as:
Serin O, Akbasli IT, Cetin SB, Koseoglu B, Deveci AF, Ugur MZ, Ozsurekci Y
Authors' Response to Peer Reviews of "Predicting Escalation of Care for Childhood Pneumonia Using Machine Learning: Retrospective Analysis and Model Development"
JMIRx Med 2025;6:e71098
URL: https://xmed.jmir.org/2025/1/e71098
doi:10.2196/71098

Authors' Response to Peer Reviews of "Data Obfuscation Through Latent Space Projection for Privacy-Preserving AI Governance: Case Studies in Medical Diagnosis and Finance Fraud Detection"

Mahesh Vaijainthymala Krishnamoorthy, BE

Stelmith, LLC, 2333 Aberdeen Pl, Carollton, TX, United States

Corresponding Author:
Mahesh Vaijainthymala Krishnamoorthy, BE
Stelmith, LLC, 2333 Aberdeen Pl, Carollton, TX, United States

Related Articles:

Companion article: https://arxiv.org/abs/2410.17459v1

Companion article: https://med.jmirx.org/2025/1/e72523

Companion article: https://med.jmirx.org/2025/1/e72525

Companion article: https://med.jmirx.org/2025/1/e70100

(JMIRx Med 2025;6:e72527) doi:10.2196/72527

KEYWORDS

privacy-preserving AI; latent space projection; data obfuscation; AI governance; machine learning privacy; differential privacy; k-anonymity; HIPAA; GDPR; compliance; data utility; privacy-utility trade-off; responsible AI; medical imaging privacy; secure data sharing; LSP; artificial intelligence

This is the authors' response to peer-review reports for "Data Obfuscation Through Latent Space Projection for Privacy-Preserving AI Governance: Case Studies in Medical Diagnosis and Finance Fraud Detection."

Round 1 Review

Reviewer AP [1]

Specific Comments

Major Comments

1. What was the basis of taking up health care cancer diagnosis and financial fraud for the study [2]? Will latent space projection be an effective method for privacy protection in speech therapy to analyze audio datasets to assist in diagnosing and treating speech-related disorders; in medical imaging video datasets from endoscopy, ultrasounds, and robotic surgeries for diagnostics and artificial intelligence (AI)–assisted tools; and in telemedicine to analyze video feeds for remote consultations and diagnoses?

Response: The basis for taking this up is to show data privacy through images and records for individuals. I would love to extend the research and will work on another paper for your suggestions. Thanks for the suggestion.

2. The basic structure of the paper is missing. Please follow the guidelines of journal paper writing with distinctly visible sections of Introduction, Method, Result/Findings, Discussion, and Limitations with future scope and conclusion. The introduction, background, and related work should be written cohesively, and all should come under the Introduction heading.

Response: I have revised the paper with major formatting changes and made it follow the Introduction-Methods-Results-Discussion formatting style as per the suggestion.

3. The statistical tables are in excess. The tables and values should be talked about in written form. Limit the number of images and tables to 5‐6 or according to the journal guidelines. Use an appendix for the flowchart and any other tabular data that is too lengthy.

Response: Statistical tables were reduced to only 3, and Figures are limited to 6 in total, but the flowchart is necessary inside the main paper.

4. Explanations of tables and figures should be in paragraph form. Please cite literature where comparative inference and process-specific benefits and drawbacks are mentioned. Examples are Tables 1-5. For writing sections like "Comparative Analysis with Existing Techniques," all the subparts should be written in paragraphs and discuss the values and analysis only, and put them in their respective paragraphs, removing the tabular data. Please use appendices for excessive tables. Within the body of the research paper, 5‐6 figures and tables are sufficient; the rest should be put in appendices.

Response: Tables have been removed and converted into paragraphs

5. In "Latency and Performance analysis, part A" and "Performance optimization" are mentions of the literature, which should be present as part of the literature in the Introduction paragraph. Restating the literature again is redundant. Stick to the structure of the journal paper. Please cite references to support the claims, such as "real-time requirements of financial systems" under the section of Real-Time Performance.

Response: Thanks; moved to the Literature section and removed from there.

6. "Scalability analysis" and other sections: What were the criteria for the choice of datasets for the study for the case studies? What were the data sizes? Give specifications in the first paragraph of respective case studies. Presenting the details about the process of procurement of files, data extraction, limitations in data handling, etc. Are there any limitations in adopting the latent space projection methods?

Response: Scalability analysis was added with the source of the dataset and the data extraction and limitations. Mostly, there are a lot of advantages compared to other privacy-preserving techniques in latent space projection; the comparative analysis proves that, and a few limitations were added as well.

Reviewer AR [3]

General Comments

I thoroughly enjoyed reading this paper as it is a well-written article that will make an important contribution to the literature on the development of privacy-preserving AI governance. I have attached a few comments to improve the study.

Response: Thanks for the compliment. Thanks for your time.

Specific Comments

Major Comments

Something like a discussion that embeds the latent space projection for AI governance and the results in the current scientific debate is missing before or after Chapter VII.

Minor Comments

In Chapter II B (Existing privacy-preserving techniques), please provide some further sources to demonstrate that the challenges mentioned are still relevant, as some sources are relatively old (eg, from 2009).

Response: I tried to address all your comments.

Round 2 Review

Reviewer AP

General Comments

This paper is highly relevant to health care, particularly in the context of privacy management of data during the analysis of imagery.

Response: Thanks for your time and effort. I appreciate it. Your comments were valuable. I addressed all your comments in this revision.

Specific Comments

Major Comments

1. The case studies should be written in a more descriptive style. Please reduce the use of numbered or bullet points (in the Introduction, Method, and Result) to align with the formal writing style typically suitable for journal papers.

Response: Removed all the bullets and converted most of them into paragraphs; some were aligned as paragraphs, but the bullet and numbered points were removed. The paper is in the Introduction-Methods-Results-Discussion format.

2. Please rephrase the description of Table 3 (immediately following the table) in a narrative style. This approach enhances the readability of the article.

Response: Rephrased the description for all the tables and figures, added descriptions for two other figures, explaining the figures deeply to make it more even, uniform, and readable, and for smooth flow.

3. Two figures should not be positioned consecutively. Include some text between Figure 3 and Figure 4. Adjust and reorganize the content to ensure a smooth flow.

Response: Addressed by adding content between 2 figures; now it makes it more readable and flows smoothly. Thanks.

Minor Comments

4. The titles of tables and figures should be presented as captions. Revise the captions to ensure they do not begin with a verb.

Response: Revised all the captions for tables and figures and made them capitalized and more readable.

Thanks for your comments.

References

1. Singh R. Peer review of "Data Obfuscation Through Latent Space Projection for Privacy-Preserving AI Governance: Case Studies in Medical Diagnosis and Finance Fraud Detection". JMIRx Med 2025;6:e72523. [doi: 10.2196/72523]
2. Vaijainthymala Krishnamoorthy M. Data Obfuscation Through Latent Space Projection for Privacy-Preserving AI Governance: Case Studies in Medical Diagnosis and Finance Fraud Detection. JMIRx Med 2025;6:e70100. [doi: 10.2196/70100]
3. Bommhardt T. Peer review of "Data Obfuscation Through Latent Space Projection for Privacy-Preserving AI Governance: Case Studies in Medical Diagnosis and Finance Fraud Detection". JMIRx Med 2025;6:e72525. [doi: 10.2196/72525]

Abbreviations

AI: artificial intelligence

Edited by CN Hang; submitted 11.02.25; this is a non–peer-reviewed article;accepted 11.02.25; published 12.03.25.

Please cite as:
Vaijainthymala Krishnamoorthy M
Authors' Response to Peer Reviews of "Data Obfuscation Through Latent Space Projection for Privacy-Preserving AI Governance: Case Studies in Medical Diagnosis and Finance Fraud Detection"
JMIRx Med 2025;6:e72527
URL: https://xmed.jmir.org/2025/1/e72527
doi:10.2196/72527

Authors' Response to Peer Reviews of "Large Language Models for Pediatric Differential Diagnoses in Rural Health Care: Multicenter Retrospective Cohort Study Comparing GPT-3 With Pediatrician Performance"

Masab Mansoor[1], BS, MBA, DBA; Andrew F Ibrahim[2], BS; David Grindem[3], DO; Asad Baig[4], MD

[1]Edward Via College of Osteopathic Medicine, 4408 Bon Aire Dr, Monroe, LA, United States

[2]Texas Tech University Health Sciences Center School of Medicine, Lubbock, TX, United States

[3]Mayo Clinic, Rochester, MN, United States

[4]Department of Radiology, Columbia University Medical Center, New York, NY, United States

Corresponding Author:
Masab Mansoor, BS, MBA, DBA
Edward Via College of Osteopathic Medicine, 4408 Bon Aire Dr, Monroe, LA, United States

Related Articles:

Companion article: https://www.medrxiv.org/content/10.1101/2024.08.09.24311777v1

Companion article: https://med.jmirx.org/2025/1/e73264

Companion article: https://med.jmirx.org/2025/1/e65263

(JMIRx Med 2025;6:e73258) doi:10.2196/73258

KEYWORDS

natural language processing; NLP; machine learning; ML; artificial intelligence; language model; large language model; LLM; generative pretrained transformer; GPT; pediatrics

This is the authors' response to peer-review reports for "Large Language Models for Pediatric Differential Diagnoses in Rural Health Care: Multicenter Retrospective Cohort Study Comparing GPT-3 With Pediatrician Performance."

We thank the reviewers [1] for the thoughtful and constructive feedback on our manuscript, "Large Language Models for Pediatric Differential Diagnoses in Rural Health Care: Multicenter Retrospective Cohort Study Comparing GPT-3 With Pediatrician Performance" [2]. We are grateful for the opportunity to revise and improve our work based on the insightful comments provided. Below, we provide detailed responses to the reviewers' comments and outline the changes made to the manuscript.

Comments and Responses

- *Please clarify why GPT-3.5 or GPT-4 (instead of GPT-3) was not used despite being available at the time of the study.*

Response: Thank you for highlighting this point. We have clarified that GPT-3 (DaVinci version) was selected because it was the most advanced version available during the study period. The Discussion section now also highlights the potential benefits of GPT-3.5 and GPT-4 for future studies, particularly in addressing rare or complex diagnoses.

Action taken: Added a rationale for GPT-3 selection in the Methods (Model Training and Fine-Tuning) section and expanded on the potential of GPT-3.5 and GPT-4 in the Discussion (GPT-3 vs Newer Models) section.

- *Why were racial and ethnic demographics not included? ("Data distribution gaps: No comparison of racial identity distribution between training and testing sets. Please consider adding a table or section on these demographic comparisons to ensure representation across subgroups.")*

Response: We acknowledge this limitation and have added a justification for the absence of this data. Specifically, the dataset lacked structured fields for racial or ethnic demographics due to its retrospective nature. We recommend future studies prioritize collecting this information to assess potential biases and ensure equitable performance.

Action taken: Added this explanation in the Materials and Methods (Participants and Data Collection) section.

- *Evaluation metrics: The study primarily uses specificity and sensitivity for evaluating large language model–generated responses, which may not capture the full quality of the outputs. Incorporating natural language processing metrics such as Recall-Oriented Understudy for Gisting Evaluation (ROUGE) and bilingual evaluation*

understudy (BLEU) can help assess the quality of generated responses more comprehensively. ROUGE measures the correspondence between the automatically generated response versus that of the human and what was expected. There are also issues associated with large language model generations of responses such as hallucination and the lack of attribution. Please specify or comment on how those and other issues were measured.

Response: We have included a discussion on hallucinations—where models generate inaccurate or unsupported outputs—and their implications for clinical use. Suggestions for addressing these issues, including the use of natural language processing metrics (eg, ROUGE and BLEU) and physician feedback mechanisms, have been added to the Discussion (Practical Implications) and Future Directions sections.

Action taken: Added text addressing hallucinations and quality evaluation in the relevant sections.

- *Figure 1 is mentioned but not included in the article, which affects comprehension of the study design and findings. Please include Figure 1 or provide an alternative reference to explain the content of the missing figure. Figures are helpful for readers to quickly grasp complex methodologies and findings.*

Response: Thank you for this suggestion. We have created and included a flowchart (Figure 1) summarizing the study workflow, including data collection, preprocessing, training/testing split, model fine-tuning, and evaluation steps.

Action taken: Added Figure 1 to the manuscript and referenced it in the appropriate sections.

- *Lack of clarity on potential implementation in rural health care settings: The study could be strengthened by detailing how the artificial intelligence (AI) model might be implemented in rural health care settings, including the specific challenges involved. Key considerations include the need for sufficient infrastructure (eg, electricity, internet) and the necessity of training health care providers unfamiliar with AI tools. Additionally, discussing both the potential impact (eg, improved diagnostic efficiency) and limitations (eg, handling incomplete data or overreliance on AI) would provide a more comprehensive road map for deployment in rural environments.*

Response: We have elaborated on the challenges of implementing AI tools in rural health care, including infrastructure limitations (eg, internet access, power supply) and costs. Recommendations for subsidized programs and partnerships with technology providers have been added to address these barriers.

Action taken: Expanded the Discussion (Practical Implications) section.

- *Address the lower accuracy for rare diagnoses.*

Response: We agree with this observation and have emphasized the need for targeted fine-tuning using domain-specific datasets to improve performance on rare pediatric conditions. This point is now discussed in the Discussion (Rare Diagnoses) section.

Action taken: Added text on targeted fine-tuning for rare diagnoses.

- *Normality test: The study does not address whether data normality was assessed before statistical analysis. Determining the distribution of the data is key to selecting the appropriate statistical test to analyze such data. The Kolmogorov-Smirnov test could aid in understanding data distribution, specifically testing for normality. If the data is not found to meet normality criteria, nonparametric methods should be applied. Including a data normality assessment and explaining the choice of a particular statistical test would significantly strengthen the reliability of the study.*

Response: Added data normality assessment details to Statistical Analysis section, specifying Kolmogorov-Smirnov testing and justification for parametric methods.

- *Power analysis assumptions: The assumptions underlying the power analysis are unclear, particularly regarding how specific diagnoses affect this analysis. It is advised to elaborate on the power analysis methodology, including the rationale behind sample size choices and their implications for diagnosis variability.*

Response: Expanded power analysis methodology with sample size rationale and considerations for diagnosis variability.

- *Sample size and generalizability: The sample size of 500 encounters may not adequately represent the broader pediatric population, particularly in diverse settings. Furthermore, using data from a single health care organization limits the applicability of findings to other settings. These limitations should be discussed, particularly how the validity of the results might change when it is tested with data from other health care centers. If possible, authors should mention and cite studies that reported on this effect. Additionally, future studies should consider expanding the sample size through multicenter collaborations or including data from patients with more diverse demographics to validate results across different health care environments thereby enhancing generalizability.*

Response: Enhanced discussion of sample size limitations with specific references to performance decreases across datasets (5%-15%).

- *Cross-validation across organizations: The model's reproducibility across various health care settings is not demonstrated. Evidence shows models often underperform with data from different sources. Including cross-organization validation and clearly acknowledging this limitation in the Discussion by citing relevant studies would enhance robustness. Furthermore, addressing this limitation in future work could pave the way for broader adoption and application of the model.*

Response: Added detailed Cross-Validation Limitations section citing studies showing model performance drops (12%-20%) across organizations.

- *Diagnostic exclusion or inclusion clarification: The preprocessing section does not clarify if physician diagnostics were included or excluded, leading to potential confusion for readers and impacting reproducibility. It would be helpful to know whether physician diagnostics were included in training and why. Clarifying this aspect would help standardize study replication and improve the study's transparency.*

Response: Clarified that physician-generated diagnoses were from retrospective data, not prospectively collected.

- *Data and model specifics for replicability: The study would benefit from more thorough descriptions of dataset characteristics, fine-tuning model parameters, and preprocessing methods. For validation, consider adding multicenter dataset details. Adding this information would enable other researchers to replicate and build upon the study's findings, thereby enhancing its scientific contribution.*

Response: Added comprehensive technical appendix with model specifications and implementation details.

- *Software and tools documentation: The authors describe using both Python (with scikit-learn) and IBM SPSS Statistics, but it is unclear what the software's sources are. Specifying sources for Python and scikit-learn (eg, "Python 3.8 [Python Software Foundation, Delaware, USA]") and clarifying the respective roles of Python and SPSS in the analyses would enhance transparency and allow for the reproducibility of the study.*

Response: Expanded Statistical Analysis section with rationale for test selection and metrics.

Additional Revisions

- Included a detailed Table 1 legend to clarify evaluation metrics (eg, true positive, false positive, true negative, and false negative).
- Added a sentence in the Future Directions section emphasizing the need for training programs tailored to rural health care providers.
- Corrected minor typographical errors in tables and sections for clarity.
- Expanded Introduction with relevant literature on large language models in pediatric contexts, including recent studies by Ramesh, Ghosh, and Haddad.

We hope these revisions address the reviewers' comments and improve the clarity, transparency, and quality of the manuscript. We sincerely thank the reviewers and the editorial team for their valuable feedback. Please do not hesitate to contact us with any additional comments or concerns.

References

1. Saderi D, Bender G, Olatoye T, et al. Peer review of "Large Language Models for Pediatric Differential Diagnoses in Rural Health Care: Multicenter Retrospective Cohort Study Comparing GPT-3 With Pediatrician Performance". JMIRx Med 2025;6:e73264. [doi: 10.2196/73264]
2. Mansoor M, Ibrahim AF, Grindem D, Baig A. Large language models for pediatric differential diagnoses in rural health care: multicenter retrospective cohort study comparing GPT-3 with pediatrician performance. JMIRx Med 2025;6:e65263. [doi: 10.2196/65263]

Abbreviations

AI: artificial intelligence
BLEU: bilingual evaluation understudy
ROUGE: Recall-Oriented Understudy for Gisting Evaluation

Edited by A Schwartz; submitted 28.02.25; this is a non–peer-reviewed article;accepted 28.02.25; published 19.03.25.

Please cite as:
Mansoor M, Ibrahim AF, Grindem D, Baig A
Authors' Response to Peer Reviews of "Large Language Models for Pediatric Differential Diagnoses in Rural Health Care: Multicenter Retrospective Cohort Study Comparing GPT-3 With Pediatrician Performance"
JMIRx Med 2025;6:e73258
URL: https://xmed.jmir.org/2025/1/e73258
doi:10.2196/73258

Authors' Response to Peer Reviews of "Improved Alzheimer Disease Diagnosis With a Machine Learning Approach and Neuroimaging: Case Study Development"

Lilia Lazli, PhD

Department of Computer and Software Engineering, Polytechnique Montréal, University of Montreal, 2500 Chem de Polytechnique, Montreal, QC, Canada

Corresponding Author:
Lilia Lazli, PhD
Department of Computer and Software Engineering, Polytechnique Montréal, University of Montreal, 2500 Chem de Polytechnique, Montreal, QC, Canada

Related Articles:

Companion article: https://arxiv.org/abs/2405.09553v1

Companion article: https://med.jmirx.org/2025/1/e73768

Companion article: https://med.jmirx.org/2025/1/e73454

Companion article: https://med.jmirx.org/2025/1/e73130

Companion article: https://med.jmirx.org/2025/1/e60866

(JMIRx Med 2025;6:e72821) doi:10.2196/72821

KEYWORDS

Alzheimer disease; computer-aided diagnosis system; machine learning; principal component analysis; linear discriminant analysis; t-distributed stochastic neighbor embedding; feedforward neural network; vision transformer architecture; support vector machines; magnetic resonance imaging; positron emission tomography imaging; Open Access Series of Imaging Studies; Alzheimer's Disease Neuroimaging Initiative; OASIS; ADNI

This is the authors' response to peer-review reports for "Improved Alzheimer Disease Diagnosis With a Machine Learning Approach and Neuroimaging: Case Study Development."

Round 1 Review

Anonymous [1]

General Comments

This paper [2] proposes a computer-aided diagnosis (CAD) system for Alzheimer disease (AD) using principal component analysis (PCA) and machine learning–based approaches. The authors claim that their system, which combines PCA for feature extraction with support vector machines (SVMs) and artificial neural networks (ANNs) for classification, achieves good accuracy in detecting AD from magnetic resonance imaging (MRI) and positron emission tomography (PET) images. However, the paper could be strengthened by addressing several areas for improvement.

Specific Comments

Major Comments

1. Consideration of alternative methodologies: While the use of PCA, SVMs, and ANNs for AD classification is a valid approach, the authors should consider exploring more recent deep learning architectures, such as vision transformers (ViTs), which have demonstrated state-of-the-art performance in medical image analysis. This would help to situate the work within the broader context of current research in the field.

Response: Done, please see the Transformers subsection (page 5). The results obtained and the discussion on the potential of this approach are mentioned in the Results (page 7) and Discussion (page 8) sections, respectively. Moreover, details on the mathematical background can be found in Multimedia Appendix 4: Vision transformer.

2. Limited evaluation: The evaluation is limited to the Open Access Series of Imaging Studies (OASIS) dataset, which may not be representative of the diverse AD population. The authors should evaluate their system on larger and more diverse datasets, such as the Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset, to demonstrate its generalizability.

Response: Done, experiments were achieved by applying the ADNI database. Please see the ADNI Data Set subsection (page 3) for more details on this basis. Table 1 (page 6) and Table 2 (page 7) for demographic characteristics and clinical assessments as well as the Results (page 7) and Discussion (page 8) sections.

Minor Comments

1. Insufficient implementation details: The implementation details of the SVMs and ANNs are insufficient. The authors should specify the hyperparameters used, such as the kernel type and regularization parameters for SVMs, and the number of layers and neurons for ANNs.

Response: Done, please see Table 3 (page 7).

2. Limited discussion: The discussion of the results is limited. The authors should provide a more in-depth analysis of the performance of their system, comparing it with other state-of-the-art methods and discussing the limitations and potential future directions.

Response: Done, please see the Discussion section (page 8).

3. The authors should ensure consistent formatting throughout the paper, including the use of italics for variables and proper capitalization in section headings.

Response: Done, the format of the journal was generally respected.

4. The paper could be improved by using more precise language. For instance, instead of "good accuracy," the authors could specify the exact accuracy percentage achieved by their system.

Response: Done, precisions for the decimal values of the results obtained are mentioned in the Results and Discussion sections, and in the abstract.

Reviewer AS [3]

General Comments

This paper explores the use of PCA and machine learning approaches for the diagnosis of AD using MRI and PET images from the OASIS database. The authors propose a system that combines PCA for feature extraction with ANNs and SVMs for classification. The paper is well structured and presents a clear methodology, but there are several areas where improvements are needed to enhance the rigor and impact of the research.

Specific Comments

Major Comments

1. Methodology justification: The choice of PCA as the sole feature extraction method needs further justification. While PCA effectively reduces dimensionality, it might not capture the most discriminative features of AD. Comparing PCA with other dimensionality reduction techniques like linear discriminant analysis or t-distributed stochastic neighbor emulation could provide a more comprehensive understanding of its effectiveness.

Response: Done, a comparative study was performed between these three dimensionality reduction techniques. Please see Table 4 and Table 5 (page 9) for the results and the Discussion section (page 8, especially lines 29-36).

2. Evaluation metrics: The paper primarily focuses on accuracy as the evaluation metric. For medical diagnosis systems, metrics like sensitivity, specificity, precision, recall, and F_1-score are crucial as they provide a better understanding of the model's performance, especially in imbalanced datasets. Including these metrics would strengthen the evaluation section.

Response: Done, please see the Statistical Analysis subsection (page 5) and Tables 4 and 5 for the results.

3. Dataset and preprocessing: The preprocessing steps are briefly mentioned but lack detailed explanation. Specific steps for noise reduction, intensity normalization, and any augmentation techniques used should be clearly described. Additionally, the impact of these preprocessing steps on the model's performance should be discussed.

Response: Done, please see the Data Preparation section (page 3) and the Discussion section (page 8), particularly, the paragraphs in lines 21 and 22 and lines 45-48.

4. Comparison with existing methods: The paper lacks a thorough comparison with existing state-of-the-art methods. Including a detailed comparison with recent literature, both in terms of methodology and performance, would provide better context and highlight the novelty and effectiveness of the proposed approach.

Response: Done, please see the Comparison With Prior Work subsection (page 9) and Table 6 (page 10).

Minor Comments

1. Introduction section: The Introduction provides a good overview of AD and the need for early diagnosis. However, it could benefit from a more detailed discussion of the current challenges in AD diagnosis and how the proposed method aims to address these challenges.

Response: The content of the Introduction has been improved to take some challenges into consideration. Please see particularly the paragraph on page 2, lines 34-48.

2. Figure and table clarity: Figures and tables should be more clearly labeled and described. For example, in Table 1, it is unclear what "Total cost (Validation)" refers to. Additionally, the axes and legends in figures should be more descriptive to enhance readability.

Response: All the content of the paper has been revised and improved by inserting new tables to clearly express the results obtained with the quantitative metrics, suggested by the evaluators. Please see the tables for the detailed results. Furthermore, the results are mentioned in the Results and Discussion sections.

3. Algorithm parameters: The specific parameters used for the SVMs and ANNs (eg, kernel type for SVMs, number of layers, and neurons for ANNs) should be explicitly mentioned. This would help in reproducing the results and understanding the model configuration.

Response: Done, please see Table 3 (page 7).

4. Conclusion and future work: The conclusion should be concise and focus on key findings. The Future Work section could be expanded to include more specific directions for further research, such as exploring different feature extraction methods, incorporating longitudinal data, or integrating other imaging modalities.

Response: This section has been deleted and replaced with the Discussion section (page 7) in order to respect the format of the journal. In this section, several subsections were inserted with content responding to your suggestion such as Main Finding (page 8) and Limitations and Future Directions (page 14).

5. References: Ensure all references are up-to-date and relevant. Given the rapid advancements in machine learning and medical imaging, some references are slightly outdated. Including more recent studies would enhance the credibility and relevance of the paper.

Response: Done, please see the references highlighted in yellow.

Anonymous [4]

General Comments

The paper discusses the development of a machine learning–based CAD system for the detection and classification of AD. The system uses brain MRI and PET images from the OASIS database, applying PCA for feature extraction and using SVMs and ANNs as classifiers. Although the proposed model shows relatively good performance, the paper should focus on justifying the novelty of the method and providing more details in the results.

Specific Comments

Major Comments

1. The paper lacks a clear discussion on how the proposed method substantially advances the state of the art. While it combines PCA with SVM and ANN, similar combinations have been explored in prior research. The authors should clearly write about how their work is novel and the specific contributions made beyond existing studies.

Response: Please see page 2, lines 34-47.

2. The paper does not provide sufficient details on the hyperparameter tuning process for both SVM and ANN models. The review suggests that the author include these additional details in an appendix.

Response: Done, Table 3 provides the hyperparameter tuning and classifiers configuration used in the experiment.

3. The evaluation primarily focuses on accuracy, sensitivity, and specificity. However, other metrics like precision, F_1-score, and area under the receiver operating characteristic curve could provide a more comprehensive assessment of the model's performance. The authors could consider adding additional metrics for evaluation.

Response: Done, other metrics were also used. Please see the Statistical Analysis section (page 5) and Table 4 and Table 5 (page 9) for the obtained results.

4. In Figure 2, the size of the box on the left and right are different (square vs rectangle). Is there a specific reason the author made this design choice?

Response: The figure was removed as more empirical results were inserted responding to the reviewers' suggestions. Techniques for reducing dimensionality and classification have been added as well as the ADNI database, which has condensed the Results and Discussion sections. I thought it wise to remove certain figures and tables to lighten the paper and avoid redundancy. However, for the design, there is no particular reason. The interface was developed using Matlab toolbox while respecting certain dimensions.

Minor Comments

1. The paper's organization can be improved. Some sections, like the methodological explanation of PCA, are overly detailed, while others, like the description of SVM and ANN, are relatively brief. Please consider balancing the sections.

Response: Done, all the content of the paper has been revised and improved. Also, appendixes were added to move the entire mathematical background and lighten the paper. Please see the Machine Learning Approaches section (page 3).

2. The Related Work section is somewhat sparse and does not sufficiently cover recent advances in the field. Please consider adding more recent studies.

Response: Done, please see the Introduction section (page 2), particularly, the paragraph in lines 21-31.

Round 2 Review

Anonymous [1]

General Comments

This paper investigates the performance of various machine learning models in the diagnosis of AD using neuroimaging data. The authors propose a CAD system that uses PCA for feature extraction and SVMs, feedforward neural networks, and ViTs for classification. The models are trained and evaluated on two datasets, OASIS and ADNI.

Specific Comments

Major Comments

1. The paper claims that the proposed CAD system is effective in classifying patients with AD and healthy controls (HCs) with high accuracy. However, the reported accuracies of 91.9% for OASIS and 88.6% for ADNI using PCA/SVM are not significantly higher than those achieved by existing state-of-the-art methods (eg, Li Y, Chen G, Wang G, et al. Dominating Alzheimer's disease diagnosis with deep learning on sMRI and DTI-MD. Front Neurol. Aug 15, 2024;15:1444795. [doi: 10.3389/fneur.2024.1444795] [PMID: 39211812]). A more comprehensive literature review and comparison are needed to support the claim of the proposed system's superiority.

Response: Performance comparisons between different machine learning techniques by referring to other researchers' studies are difficult. It is possible that the same algorithm can provide different results for the same database if the study context, the

acquisition and learning parameters, the capacity of the computing equipment, etc are different. Nevertheless, to evaluate the effectiveness of the proposed CAD system, a comparative study with some recent works was carried out on the ADNI and OASIS datasets, which we think the development conditions are almost similar to our case.

An objective comparison could not be made with the study proposed in the *Frontiers in Neurology* paper you suggested for two reasons.

1. Researchers used samples from a mixture of two databases, ADNI and Xuanwu Hospital Neuroimaging, to perform the training of the CNN. This provides more data to conduct this process well.
2. Researchers performed two binary classifications (AD vs HCs and mild cognitive impairment [MCI] vs HCs), and they obtained accuracies of 0.96% and 0.83% respectively. In our case, the binary classification performed is AD vs HCs, where samples from patients with MCI and those with confirmed AD are grouped in the same Alzheimer class. The ViT model achieved an accuracy of 90.4% for this category, which is encouraging because MCI is a difficult stage to predict.

2. The ADNI dataset includes not only patients with AD and HCs but also individuals with MCI. The paper does not explicitly mention whether MCI cases are included in the ADNI dataset used in this study and if patients with MCI are excluded. What is the reason?

Response: Clarifications are provided regarding the subdivision of the two HC and AD classes, which concern HCs and patients with AD, respectively. Please see the related paragraphs on page 3.

3. The paper's conclusion that the "PCA/SVM scheme is much better at predicting AD than the other models" is not supported by the results presented. The ViT model with data augmentation consistently outperforms PCA/SVM in terms of accuracy and other metrics. There are no obvious reasons data augmentation is unwanted either.

Response: Details are provided regarding the results obtained with the ViT classifier. Please see the related paragraphs on page 1 and page 2 in the abstract section.

We have confirmed your deduction regarding the performance of the ViT that was applied in conjunction with the data augmentation strategy. We have not criticized the potential of having augmented the data. In general, neural networks in comparison with other machine learning models need a sufficient amount of data to perform their training in order to obtain good results. Therefore, in cases with little data, it is necessary to go through strategies that allow increased data to achieve this objective.

In the paragraph titled Method in the abstract section, we have specified that three classifiers were used: SVM and FFNN with the dimensionality reduction methods as well as ViT with the data augmentation strategy. The Results and Conclusion subsections in the abstract section confirmed that the data augmentation/ViT model outperformed the other models.

Minor Comments

1. The paper claims to use a multimodal system, combining both MRI and PET images. However, it does not compare the multimodal system's performance against single-modal systems using only MRI or PET images. Such a comparison would help to rationalize the conclusion that the multimodal system truly improves upon single-modal systems.

Response: Please see the related paragraph on page 8.

Reviewer AS

General Comments

Thank you for addressing my comments from the previous round of reviews. I appreciate the effort you have put into revising the manuscript. The updated version effectively resolves all the issues I raised, and the manuscript is now clear, well-structured, and scientifically sound.

Response: Thank you very much for your valued contribution as well as for your relevant comments in round 1, which helped to improve the contents of the paper.

References

1. Anonymous. Peer review of "Improved Alzheimer Disease Diagnosis With a Machine Learning Approach and Neuroimaging: Case Study Development". JMIRx Med 2025;6:e73768. [doi: 10.2196/73768]
2. Lazli L. Improved Alzheimer Disease Diagnosis With a Machine Learning Approach and Neuroimaging: Case Study Development. JMIRx Med 2025;6:e60866. [doi: 10.2196/60866]
3. Khani M. Peer review of "Improved Alzheimer Disease Diagnosis With a Machine Learning Approach and Neuroimaging: Case Study Development". JMIRx Med 2025;6:e73454. [doi: 10.2196/73454]
4. Anonymous. Peer review for "Improved Alzheimer Disease Diagnosis With a Machine Learning Approach and Neuroimaging: Case Study Development". JMIRx Med 2025;6:e73130. [doi: 10.2196/73130]

Abbreviations

AD: Alzheimer disease
ADNI: Alzheimer's Disease Neuroimaging Initiative
ANN: artificial neural network
CAD: computer-aided diagnosis

HC: healthy control
MCI: mild cognitive impairment
MRI: magnetic resonance imaging
OASIS: Open Access Series of Imaging Studies
PCA: principal component analysis
PET: positron emission tomography
SVM: support vector machine
ViT: vision transformer

Edited by CN Hang; submitted 18.02.25; this is a non–peer-reviewed article;accepted 18.02.25; published 21.04.25.

Please cite as:
Lazli L
Authors' Response to Peer Reviews of "Improved Alzheimer Disease Diagnosis With a Machine Learning Approach and Neuroimaging: Case Study Development"
JMIRx Med 2025;6:e72821
URL: https://xmed.jmir.org/2025/1/e72821
doi:10.2196/72821

Authors' Response to Peer Reviews of "Levels and Predictors of Knowledge, Attitudes, and Practices Regarding Contraception Among Female TV Studies Undergraduates in Nigeria: Cross-Sectional Study"

Hadizah Abigail Agbo[1,2*], MBBS, MPH, MSc, FWACP; Philip Adewale Adeoye[1*], MBBS, MWACP, MPH; Danjuma Ropzak Yilzung[3*], MBBS; Jawa Samson Mangut[3*], MBBS; Paul Friday Ogbada[3*], MBBS

[1]Department of Community Medicine, Jos University Teaching Hospital, Lamingo, Jos, Plateau State, Nigeria

[2]Department of Community Medicine, University of Jos, Jos, Plateau State, Nigeria

[3]College of Health Sciences, University of Jos, Jos, Plateau State, Nigeria

[*]all authors contributed equally

Corresponding Author:
Philip Adewale Adeoye, MBBS, MWACP, MPH
Department of Community Medicine, Jos University Teaching Hospital, Lamingo, Jos, Plateau State, Nigeria

Related Articles:

Companion article: https://www.medrxiv.org/content/10.1101/2024.01.01.24300698v1

Companion article: https://med.jmirx.org/2025/1/e72951

Companion article: https://med.jmirx.org/2025/1/e72949

Companion article: https://med.jmirx.org/2025/1/e56135

(JMIRx Med 2025;6:e72947) doi:10.2196/72947

KEYWORDS

knowledge; attitudes; practice; contraception; regression; cross-sectional; females; students; Nigeria

This is the authors' response to peer-review reports of "Levels and Predictors of Knowledge, Attitudes, and Practices Regarding Contraception Among Female TV Studies Undergraduates in Nigeria: Cross-Sectional Study."

Round 1 Review

Reviewer Q [1]

General Comments

Dear Authors,

Thank you very much for undertaking the study [2] titled "Levels and predictors of knowledge, attitude and practice of contraception among female TV undergraduates in Nigeria: a cross-sectional study" and submitting the manuscript to JMIR. The study findings are important for family planning program implementation targeting young students. I have the following comments and observations for improving your manuscript for consideration of publishing.

Response: We would like to thank the reviewer for the kind words and helpful comments. We are indeed grateful.

Specific Comments

Major Comments

Introduction: line 50: "youth": Indicate age group.

Response: Done. The age range (17 - 35 years) of the study participants falls within the definition of youth by the National Baseline Youth Survey of Nigeria and the African Youth Charter [3,4]. The classification used for teenagers/adolescents agrees with the World Health Organization definition and those used in the literature on adolescents [5,6]. The classification of young adults used agrees with that of Statistics Canada [7].

Line 52: "Utilization is higher": Not clear what the utilization was for.

Response: Revised to "contraceptive utilization rate."

Study population: limitation: gender biased. Male involvement and attitude are equally important regarding sexually transmitted infections, particularly for male methods like use of condoms. This needs to be mentioned as a limitation of the study.

Response: Done.

Tables all: Hastily, one sentence is used for describing findings in a table. Need to elaborate more. Further comments below.

Response: Tables are now more elaborate in narration.

Table 1: Rephrase the "Marital status" indicator; the data does not give the status of marriage!

Response: The original marital status categorization on the data collection form includes single, married, separated, divorced, and widowed. However, after the data collection, only single (n=197, 96.8%), married (n=19, 8.8%), and separated (n=1, 0.5%) were reported. Since only 1 study participant reported herself as separated and this group is similar to singles by not living with their spouse, they were, therefore, merged to ease the interpretation of data and to reflect the impact of living with a spouse on contraceptive attitude and use. Therefore, I have rephrased the marital status grouping as married and single/separated.

Table 2: Indicate what is meant by poor, good, etc knowledge/attitude; cite measurement scale here.

Response: It has already been stated in the Data Management and Analysis subsection of the Methodology section. The classification is based on the use of the average scores. This is dependent on whether they are normally distributed or not: when they are normally distributed, mean (SD) was used, but when not normally distributed, median (IQR) was reported. Good knowledge, attitude, and practice are at least the average scores; while poor knowledge, attitude, and practice are less than the average scores. This approach to categorization is important to prevent the "ceiling effect" in subjective socially biased items in surveys. Therefore, the categorization scale has been indicated within the narration of the result as requested.

Table 3: Need to mention if this was an open-ended or structured question.

Response: It has already been stated under the data collection methods: "Data was collected from female students of NTA TV College Jos by the research team using a semi-structured self-administered questionnaire..." The questionnaire is semistructured and contains both structured and open-ended items.

Table 4: Cite the indicators used for measuring attitude toward use of contraception.

Response: Indicators include those items explored in the secondary analysis of this data, which has been posted on a preprint server to provide insight into the items driving the reported levels and predictors of contraception reported in this study [8].

Table 5: The predictor of not engaging in sex may be reflected well in statistical analysis, but what is the significance in real life? Why would those who had never engaged in sex have used contraception?

Response: Though it might not be relatively acceptable and valid to ask those who have never had sexual intercourse about contraceptive use, the researchers were prompted to generally ask this question due to the prevalence of intimate partner violence among unmarried and separated people with the prevalence of sexual violence being one of the highest in the country; much earlier first sexual experiences among the age range in the study population in the study area, region, and country, with first sexual experience not forced; increasing use of contraception for other purposes other than family planning among the study population; increasing liberal attitudes toward contraceptive use; and the social desirability bias that can be produced with questions surrounding sex and contraception [9-11]. We were justified by the time we explored the result of the responses and the inconsistencies reported by the study participants; some of the results were added to a preprint published earlier this year, but 73.7% of this study population reported having had sex, and a higher proportion (94.9%) of the total study population reported history of unplanned pregnancy [8,12]. That might have been the reason, among many others, why many other published studies have included the same item in questionnaires for all study participants irrespective of declared sexual activity status [13-15].

Discussion: Mention the rate of use of emergency contraceptive pills (ECPs) also. This is increasing in many societies. Policy makers/planners are often not aware of the need for ECPs to include a supply of ECPs in a program.

Response: The report to reflect the use of emergency contraception has been expanded in the Result section. Due to the small proportion of study participants using emergency contraception, they have been merged with those reporting implants and many unnamed forms of contraception in the "others category." Further discussion on emergency contraception use has been included in the Discussion also.

A recommendation like "There may be a need to use social marketing 42 approaches to make these contraceptives available to young people to bypass the stigma they experienced while accessing 43 contraceptives from traditional sources of contraceptives" is not supported by any finding or data of the study. Rather this raises a question of bias on jumping to a solution through a particular channel. Let the program planners find out the way to resolve the issue of information availability.

Response: The discussion on social marketing implications under contraceptive use has been removed following the recommendation of the reviewer. However, social marketing is a veritable tool in ensuring improved access to contraception for young people using marketing approaches. Its recognized ability to increase use also prompted its inclusion in the Nigeria Demographic and Health Survey 2018 for the first time, where women of reproductive age (15 - 49 years) in the country (including the study area) were asked about the use of social marketing to access contraception [9].

Highlights: Move the highlights to the Discussion section because this is a summary of the findings.

Response: Done.

Conclusion: Rewrite the conclusion, elaborating on recommendations per results of the study.

Response: Done. Other important recommendations have been added to the discussion of important results. Discussions usually include a statement on a result, comparison/variation with prior

studies, and reasons for the similarities/variations and implications for policy and practice.

Reviewer BO [16]

Specific Comments

Major Comments

1. The sampling technique used in this paper should be more detailed than it is. Respondents were said to have been selected by balloting from the 6 levels. Was it equal allocation per level, or was it proportionate allocation considering that it is not likely that there were the same number of students in each level?

Response: A detailed sampling has been reported in the work as requested by the reviewer.

2. State the age ranges of a teenager and that of a young adult in your methodology that informed the categorization in the Results.

Response: Done.

3. Living with a spouse and not living with a spouse was considered for marital status in your study as opposed to being single, married, etc. Clarify why this is so.

Response: The original marital status categorization on the data collection form included single, married, separated, divorced, and widowed. However, after the data collection, only single (n=197, 96.8%), married (n=19, 8.8%), and separated (n=1, 0.5%) were reported. Since only 1 study participant reported herself as separated and this group is similar to singles by not living with their spouse, they were merged to ease the interpretation of data and to reflect the impact of living with a spouse on contraceptive attitude and use. Therefore, I have rephrased the marital status grouping as married and single/separated.

4. The public health implications of some of the findings were omitted in the Discussion. This should be included. Its importance cannot be overemphasized.

Response: Done.

Minor Comments

5. Abstract: The last sentence in the Methods is hanging. Kindly complete it.

Response: Done.

6. Grammatical issues: Tenses: Future and present tenses were used where past tense should have been used in the methodology (lines 12 and 28). Present tense was used in multiple places in the Discussion where past tense should have been used.

Response: Done.

7. Reference list: In the Vancouver referencing style, the month of publication should not appear as it did in some references like 7, 11, and 12. The date accessed/cited was written in some and not in others like 9, 10, 13, and 16. Really old references like reference 24, which is 14 years old, should be replaced by more current ones.

Response: Done. “Month of publication” as seen in some journal references had been removed from the reference list.

Revised to conform to stated format. Months of access were included in websites as seen in references 16, 17, 18, 21, 22, 35, and 36; while they were not included in references 1, 4, 6, 9, 10, 13, 45, and 48 because access dates are not necessarily included in reports. Also, to prevent unnecessary errors in referencing, Mendeley referencing software was used.

Really old references (2) have been replaced [17,18]. Others (3) were left because they are either a charter or government document that contribute to a definition [3,4] or milestone document [19].

References

1. Biswas KK. Peer review of “Levels and Predictors of Knowledge, Attitudes, and Practices Regarding Contraception Among Female TV Studies Undergraduates in Nigeria: Cross-Sectional Study”. JMIRx Med 2025;6:e72949. [doi: 10.2196/72949]
2. Agbo HA, Adeoye PA, Yilzung DR, Mangut JS, Ogbada PF. Levels and predictors of knowledge, attitudes, and practices regarding contraception among female TV studies undergraduates in Nigeria: cross-sectional study. JMIRx Med 2025;6:e56135. [doi: 10.2196/56135]
3. National baseline youth survey: final report. National Bureau of Statistics. 2012. URL: https://www.nigerianstat.gov.ng/pdfuploads/2102 National Baseline Youth Survey Report_1.pdf [accessed 2025-04-01]
4. African Youth Charter. UNFPA East and Southern Africa. 2006. URL: https://esaro.unfpa.org/sites/default/files/pub-pdf/CHARTER_English.pdf [accessed 2025-04-01]
5. Adolescent health. World Health Organization. 2024. URL: https://www.who.int/health-topics/adolescent-health#tab=tab_1 [accessed 2025-04-01]
6. Liang M, Simelane S, Fortuny Fillo G, et al. The state of adolescent sexual and reproductive health. J Adolesc Health 2019 Dec;65(6S):S3-S15. [doi: 10.1016/j.jadohealth.2019.09.015] [Medline: 31761002]
7. Infographic 5: plateau in the share of young adults living with their parents from 2016 to 2021. Statistics Canada. 2022. URL: https://www150.statcan.gc.ca/n1/daily-quotidien/220713/g-a005-eng.htm [accessed 2025-04-01]
8. Adeoye PA, Adeniji T, Agbo HA. Predictors of good contraception attitude and practice among female students of television studies in Nigeria: a secondary analysis. medRxiv. Preprint posted online on Aug 4, 2024. [doi: 10.1101/2024.02.26.24303367]
9. Nigeria Demographic and Health Survey 2018. The DHS Program. 2019 Oct. URL: https://dhsprogram.com/pubs/pdf/FR359/FR359.pdf [accessed 2025-04-01]

10. Okoh E, Ismaila E, Noel B, et al. Prevalence and predictors of intimate partner violence among women of reproductive age in Plateau State, North-Central Nigeria. J Women Child Health 2024 Mar 19;1(2):21-27. [doi: 10.62807/jowach.v2i1.2024.21-27]
11. Dragoman MV. The combined oral contraceptive pill -- recent developments, risks and benefits. Best Pract Res Clin Obstet Gynaecol 2014 Aug;28(6):825-834. [doi: 10.1016/j.bpobgyn.2014.06.003] [Medline: 25028259]
12. Agbo HA, Adeoye PA, Yilzung DR, Mangut JS, Ogbada PF. Levels and predictors of knowledge, attitude and practice of contraception among female TV undergraduates in Nigeria: a cross-sectional study. Preprint posted online on Jan 2, 2024. [doi: 10.1101/2024.01.01.24300698]
13. Eniojukan JF, Ijeoma O, Prince O, et al. Knowledge, perception and practice of contraception among staff and students in a university community in Delta State, Nigeria. Pharm Biosciences J 2015;4(1):71-81. [doi: 10.20510/ukjpb/4/i1/87848]
14. Sweya MN, Msuya SE, Mahande MJ, Manongi R. Contraceptive knowledge, sexual behavior, and factors associated with contraceptive use among female undergraduate university students in Kilimanjaro region in Tanzania. Adolesc Health Med Ther 2016 Oct 3;7:109-115. [doi: 10.2147/AHMT.S108531] [Medline: 27757057]
15. Somba MJ, Mbonile M, Obure J, Mahande MJ. Sexual behaviour, contraceptive knowledge and use among female undergraduates' students of Muhimbili and Dar es Salaam Universities, Tanzania: a cross-sectional study. BMC Womens Health 2014 Aug 7;14(1):94. [doi: 10.1186/1472-6874-14-94] [Medline: 25099502]
16. Nwankwo B. Peer review of "Levels and Predictors of Knowledge, Attitudes, and Practices Regarding Contraception Among Female TV Studies Undergraduates in Nigeria: Cross-Sectional Study". JMIRx Med 2025;6:e72951. [doi: 10.2196/72951]
17. Jones RK, Biddlecom AE, Hebert L, Mellor R. Teens reflect on their sources of contraceptive information. J Adolesc Res 2011 Mar 14;26(4):423-446. [doi: 10.1177/0743558411400908]
18. Tilahun D, Assefa T, Belachew T. Knowledge, attitude and practice of emergency contraceptives among adama university female students. Ethiop J Health Sci 2010 Nov;20(3):195-202. [doi: 10.4314/ejhs.v20i3.69449] [Medline: 22434979]
19. Cohen SA. London summit puts family planning back on the agenda, offers new lease on life for millions of women and girls. Guttmacher Policy Rev 2012;15(3):20-24 [FREE Full text]

Abbreviations

ECP: emergency contraceptive pill

Edited by A Schwartz; submitted 21.02.25; this is a non–peer-reviewed article;accepted 21.02.25; published 08.05.25.

Please cite as:
Agbo HA, Adeoye PA, Yilzung DR, Mangut JS, Ogbada PF
Authors' Response to Peer Reviews of "Levels and Predictors of Knowledge, Attitudes, and Practices Regarding Contraception Among Female TV Studies Undergraduates in Nigeria: Cross-Sectional Study"
JMIRx Med 2025;6:e72947
URL: https://xmed.jmir.org/2025/1/e72947
doi:10.2196/72947

Authors' Response to Peer Reviews of "Prevalence and Determinants of Academic Bullying Among Junior Doctors in Sierra Leone: Cross-Sectional Study"

Fatima Jalloh[1], MB ChB; Ahmed Tejan Bah[2], MB ChB, MPH; Alieu Kanu[3], MB ChB; Mohamed Jan Jalloh[3], MB ChB; Kehinde Agboola[3], MB ChB; Monalisa M J Faulkner[3], MB ChB; Foray Mohamed Foray[4], MB ChB, MPH; Onome T Abiri[1], BPharm, PharmD, MSc; Arthur Sillah[5], PhD; Aiah Lebbie[1], MB ChB; Mohamed B Jalloh[6], MB ChB, MSc

[1]College of Medicine and Allied Health Sciences, University of Sierra Leone, Freetown, Sierra Leone

[2]Department of Public Health, Chamberlain College of Health Professions, Chicago, IL, United States

[3]University of Sierra Leone Teaching Hospitals Complex, Freetown, Sierra Leone

[4]College of Health Sciences and Public Policy, Walden University, Minneapolis, MN, United States

[5]School of Public Health, University of Washington, Seattle, WA, United States

[6]Faculty of Health Sciences, Department of Medicine, McMaster University, 1280 Main Street West, Hamilton, ON, Canada

Corresponding Author:
Mohamed B Jalloh, MB ChB, MSc
Faculty of Health Sciences, Department of Medicine, McMaster University, 1280 Main Street West, Hamilton, ON, Canada

Related Articles:

Companion article: https://www.medrxiv.org/content/10.1101/2024.11.13.24317261v1

Companion article: https://med.jmirx.org/2025/1/e75134

Companion article: https://med.jmirx.org/2025/1/e75135

Companion article: https://med.jmirx.org/2025/1/e68865

(JMIRx Med 2025;6:e75127) doi:10.2196/75127

KEYWORDS

academic bullying; junior doctors; Sierra Leone; mental health; professional development

This is the authors' response to peer-review reports for "Prevalence and Determinants of Academic Bullying Among Junior Doctors in Sierra Leone: Cross-Sectional Study."

Round 1 Review

Reviewer AQ [1]

Specific Comments

Major Comments

Introduction

I think the Introduction in this study [2] needs to be contextualized properly. Saying that bullying in the health care profession has not been looked at is largely correct, but the authors need to strengthen their argument by properly discussing the current literature on bullying in the Sierra Leone educational establishment and the limitations of the current literature as it relates to their topic of enquiry.

Please read the following:

- *Osborne A, James PB, Bangura C, Tom Williams SM, Kangbai JB, Lebbieie, A. Bullying victimization among in-school adolescents in Sierra Leone: a cross-sectional analysis of the 2017 Sierra Leone Global School-Based Health Survey. PLOS Glob Public Health. Dec 22, 2023;3(12):e0002498. [doi: 10.1371/journal.pgph.0002498] [PMID: 38134001]*
- *Report on findings from school-related gender-based violence action research in schools and communities in Sierra Leone [3].*

Response: We thank the reviewer for their helpful feedback and suggested references. We have revised and expanded the Introduction section with suggested references (see pages 4 and 5).

Methods

I wonder why the authors decided not to recruit all junior doctors who met their inclusion criteria, given that the list of junior doctors in the University of Sierra Leone Teaching Hospitals Complex at the time of data collection can be obtained

from each of the constituent teaching hospitals. I know for a fact that the population of junior doctors is not so huge (less than 500). In other words, why did the authors just recruit all 160 junior doctors? Such data can be sourced from the Sierra Leone Medical and Dental Association or from the respective teaching hospital.

Response: Thank you for highlighting this important point. We recognize that the total population of junior doctors at these facilities is indeed under 500. Our original intention was to recruit all eligible junior doctors, which would have strengthened the study's power and rendered sample size calculations less critical. However, achieving a 100% response rate proved difficult—particularly given the 3- to 6-month rotation schedules that complicate maintaining an up-to-date sampling frame. Consequently, we used a pragmatic sampling strategy, distributing the survey through the Sierra Leone Medical and Dental Association forums and across the respective hospitals for several weeks. While this approach did not capture every potential respondent, it yielded a sufficiently robust sample to draw meaningful conclusions despite the inevitable limitations of incomplete participation.

What informed the design of the questionnaire used? Why did the authors decide not to conduct any form of validation of the questionnaire (ie, externally or internally) to ensure it is appropriate for the context in which it is used?

Response: Thank you for your insightful questions regarding the questionnaire design and validation. Our questionnaire was primarily informed by prior studies from the subregion—most notably the work by Afolaranmi et al [4] in Nigeria, whose clinical training context is highly comparable to Sierra Leone. Given that many Sierra Leonean medical educators and clinical trainers received their training in Nigeria and a number of Nigerian professors practice in Sierra Leone, we found these instruments to be a suitable starting point.

To enhance contextual relevance, we conducted a pilot with 10 participants to assess clarity, applicability, and cultural appropriateness prior to rolling out the full study (see page 7). However, we acknowledge the lack of a psychometric validated tool in the manuscript's Limitations section.

This study was among junior doctors, but the authors mentioned registrars. A registrar is no longer a junior doctor. I may be wrong, but I strongly suggest that the authors provide a clear definition of what is the definition of junior doctor in Sierra Leone.

Response: Thank you for raising this important clarification. In many settings, the term "registrar" refers to a physician who has moved beyond the intern or house officer stage and may be considered more senior. However, in the context of Sierra Leone's postgraduate training system, registrars still fall within the broader category of early-career physicians, who have not yet obtained final specialist accreditation.

To be specific, a "junior doctor" in Sierra Leone typically includes:

- House officers/interns, who have recently graduated and are completing supervised practice
- Medical officers, who work more independently but have not pursued formal residency training
- Registrars (residents), who are enrolled in specialty training programs and have not yet become fully accredited specialists

This aligns with the general World Medical Association perspective that "junior doctors" encompass physicians in postgraduate training who have not yet achieved final specialty qualification. In Sierra Leone, this definition covers registrars, as they remain in an active training pathway and do not possess full consultant status. Hence, our study included registrars under the umbrella of "junior doctors." We hope this clarifies why registrars were incorporated into our sample.

Discussion

I beg to disagree. A sample was calculated, and a probabilistic sampling method was used in this study, which means that it gives an equal chance for everyone to be chosen. Thus, the sample used is representative of junior doctors in the University of Sierra Leone Teaching Hospitals Complex. There are two ways to explain your finding: either the sample is not representative because the sampling was not probabilistic or the whole population should have been recruited, or the finding is correct (ie, there are no gender differences).

Response: Thank you for your insightful feedback. We fully acknowledge that our study was designed with a calculated sample size and a probabilistic sampling method, with the aim of ensuring a representative sample of junior doctors in the University of Sierra Leone Teaching Hospitals Complex. This design typically affords every eligible participant an equal opportunity to be selected. Thus, our finding of no statistically significant gender differences in bullying could indeed reflect a true lack of disparity within this specific population.

We appreciate your perspective and have revised the Discussion to more clearly articulate these points.

Minor Comments

The first two sentences of the third paragraph of the Introduction section: This has already been stated in the previous paragraph. This is just a repetition.

Response: Thank you. We have revised the "Introduction" section with suggested changes (see pages 4 and 5).

Round 2 Review

Reviewer EN [5]

General Comments

This study presents a survey of junior doctors in Sierra Leone hospitals and their experience of bullying and found high levels of bullying among the participants. Below are comments and suggestions for clarifying and strengthening the work.

Specific Comments

Major Comments

1. The author's definition of bullying and whether it was provided to participants is somewhat unclear. In the abstract,

bullying is described as involving repeated behaviors, which aligns with the typical definition of bullying as an ongoing or repeated action. However, in the Methods section, participants were asked to respond based on any instance of various behaviors. While a single act of intimidation, for example, constitutes inappropriate behavior that should be addressed, it may not meet the standard definition of bullying. It is essential to clarify this distinction and ensure that participants also recognized the difference so that general poor behavior is not conflated with bullying.

Response: Thank you for emphasizing this point. Our study was conducted using the recognized definition of bullying as involving repeated behaviors. In our original design and implementation, we informed participants that bullying typically denotes a pattern of ongoing or repeated actions. We acknowledge, however, that some of our language in the manuscript may have led to confusion around single versus repeated incidents. We have therefore reviewed and refined our wording throughout the text—particularly in the abstract and Methods section—to ensure consistent use of the term "bullying" and to clarify that isolated, one-time acts, while concerning, may not meet the standard definition of repeated harmful behavior (see page 7).

2. Was sampling randomly, equally, or proportionally distributed across the four sites, and were there any analyses done based on site?

Response: Our sampling was designed to be random at the individual level rather than equally or proportionally allocated to each site. Because junior doctors rotate across the four sites at the University of Sierra Leone Teaching Hospitals Complex, we treated all eligible doctors as a single sampling frame. Each individual had an equal probability of selection through a computer-based random procedure, independent of their current site.

Regarding site-level analyses, we elected not to perform them because the frequent rotations diminished the value of comparing departments as distinct groups. Instead, we focused on the overall experiences of junior doctors within the hospital complex. Any subgroup analysis by site would have been confounded by the high degree of overlap in personnel across the four locations (see page 6).

3. How was random sampling achieved?

Response: Thank you for highlighting this important methodological detail. We ensured that each eligible junior doctor had an equal probability of being included by employing a computer-based random selection procedure. Specifically:

- Comprehensive sampling frame: We first compiled a roster of all junior doctors who met our eligibility criteria (aged ≥18 years and employed at the University of Sierra Leone Teaching Hospitals Complex for ≥6 months).
- Unique identifiers: Each individual in this roster was assigned a unique numeric code.
- Random number generation: We then used a random number generator to select participants based on their assigned numeric codes, thereby ensuring that every eligible junior doctor had the same chance of selection.

This approach was chosen to reduce selection bias and maintain methodological rigor, despite the logistical challenges posed by junior doctors' frequent rotations across departments (see page 6).

4. Please comment on the reliability and validity of the instrument used to collect data. What literature was used to inform the development of the questions? Please include this information in the manuscript.

Response: Thank you for your insightful questions regarding the questionnaire design and validation. Our questionnaire was primarily informed by prior studies from the subregion—most notably the work by Afolaranmi et al [4] in Nigeria, whose clinical training context is highly comparable to Sierra Leone. Given that many Sierra Leonean medical educators and clinical trainers received their training in Nigeria and a number of Nigerian professors practice in Sierra Leone, we found these instruments to be a suitable starting point.

To enhance contextual relevance, we conducted a pilot with 10 participants to assess clarity, applicability, and cultural appropriateness prior to rolling out the full study (see page 7). However, we acknowledge the lack of a psychometric validated tool in the manuscript's Limitations section.

5. At the start of paragraph 3 of the Introduction, the authors refer to "other contexts"; it is unclear what contexts are being referred to in this and the preceding paragraph.

Response: We thank the reviewer for their helpful feedback and suggested references. We have revised and expanded the Introduction section, including suggested references by another reviewer (see pages 4 and 5).

6. The Introduction and Discussion would be strengthened by more specific references to literature findings. I found the text in both a little superficial.

Response: We thank the reviewer for their helpful feedback. We have revised and expanded the Introduction and Discussion sections, including suggested references by another reviewer (see pages 4, 5, and 11 - 15).

7. It is unclear whether the participants were reporting behaviors they personally experienced (ie, they were bullied) against behaviors they observed (ie, others being bullied).

Response: We specifically designed our questionnaire to capture bullying events that respondents personally experienced, rather than those they witnessed. The survey items regarding workplace bullying were phrased to reflect direct, firsthand encounters. Respondents who indicated experiencing bullying were then asked to describe the nature of these incidents, ensuring the data represented self-reported victimization rather than secondhand observations (see page 7).

8. Please provide clarification as to who is a "junior doctor." This journal has an international readership, and this term can be used differently in different countries, with "junior doctors" having different lengths of service. Please ensure this is clear within the body of the manuscript.

Response: Thank you for noting this. In Sierra Leone, the term "junior doctor" encompasses three main groups:

- House officers/interns: recently graduated doctors in a period of closely supervised practice
- Medical officers: physicians who have completed internships and can work more independently but have not pursued formal residency training
- Registrars (residents): doctors actively enrolled in specialty training programs who have not yet attained full consultant (specialist) status

This aligns with the broader World Medical Association definition, which frames "junior doctors" as physicians in postgraduate training who have not yet achieved their final specialty qualifications. We have included all three categories in our study, as they each fulfill the criteria of postgraduate training without full specialist accreditation (see pages 5 and 6).

9. The description of the multiple regression seems a little excessive given the lack of statistical significance. This could be made more concise and simply refer readers to Table 3. Similarly, the authors should be cautious not to overemphasize these findings.

Response: Thank you for this valuable feedback. We appreciate the concern about potentially overstating findings that did not reach statistical significance. We believe it is important to retain the full results for completeness and transparency—even when no statistically significant associations emerge. In light of your suggestion, we will ensure that our manuscript clearly indicates the nonsignificant nature of these results and refrain from overemphasizing their importance in the Discussion.

10. The list of references needs to be reviewed to ensure that all items have full bibliographic details.

Response: Thank you for noting this. We have carefully reviewed and updated the reference list to ensure that all citations include complete bibliographic details.

Conflicts of Interest

None declared.

References

1. James PB. Peer review of "Prevalence and Determinants of Academic Bullying Among Junior Doctors in Sierra Leone: Cross-Sectional Study". JMIRx Med 2025;6:e75134. [doi: 10.2196/75134]
2. Jalloh F, Bah AT, Kanu A, et al. Prevalence and determinants of academic bullying among junior doctors in Sierra Leone: cross-sectional study. JMIRx Med 2025;6:e68865. [doi: 10.2196/68865]
3. Report on findings from school-related gender-based violence action research in schools and communities in Sierra Leone. United Nations Girls' Education Initiative. URL: https://www.ungei.org/publication/report-findings-school-related-gender-based-violence-action-research-schools-and [accessed 2025-04-16]
4. Afolaranmi TO, Hassan ZI, Gokir BM, et al. Workplace bullying and its associated factors among medical doctors in residency training in a tertiary health institution in Plateau State Nigeria. Front Public Health 2021;9:812979. [doi: 10.3389/fpubh.2021.812979] [Medline: 35155359]
5. Wilkinson J. Peer review of "Prevalence and Determinants of Academic Bullying Among Junior Doctors in Sierra Leone: Cross-Sectional Study". JMIRx Med 2025;6:e75135. [doi: 10.2196/75135]

Edited by S Tungjitviboonkun; submitted 28.03.25; this is a non–peer-reviewed article;accepted 28.03.25; published 22.05.25.

Please cite as:
Jalloh F, Bah AT, Kanu A, Jalloh MJ, Agboola K, Faulkner MMJ, Foray FM, Abiri OT, Sillah A, Lebbie A, Jalloh MB
Authors' Response to Peer Reviews of "Prevalence and Determinants of Academic Bullying Among Junior Doctors in Sierra Leone: Cross-Sectional Study"
JMIRx Med 2025;6:e75127
URL: https://xmed.jmir.org/2025/1/e75127
doi:10.2196/75127

Author's Response to a Commentary on "Prevalence of Undiagnosed Hypertension Among Adult Displaced Individuals in Baidoa Camps, Somalia (Preprint)"

Mohamed Jayte

Internal Medicine Department, Kampala International University, F46V+MW2, Ishaka, Kampala, Uganda

Corresponding Author:
Mohamed Jayte
Internal Medicine Department, Kampala International University, F46V+MW2, Ishaka, Kampala, Uganda

Related Articles:

Companion article: https://www.medrxiv.org/content/10.1101/2024.03.22.24304736v1

Companion article: https://med.jmirx.org/2025/1/e71041

(JMIRx Med 2025;6:e70265) doi:10.2196/70265

KEYWORDS

prevalence; undiagnosed; epidemiology; heart; cardiology; cardiovascular; cross-sectional; survey; questionnaires; hypertension; blood pressure; poverty; sedentary; displaced; refugee; Africa

This is the author's response to a commentary on "Prevalence of Undiagnosed Hypertension Among Adult Displaced Individuals in Baidoa Camps, Somalia."

Round 1 Review

Anonymous [1]

I commend the author for this study [2] on an important topic. However, here are a few comments to help improve the manuscript.

Title

1. The title needs some slight changes to improve clarity. For instance, what do you mean by "displaced individuals"? Would you rather state it as "internally displaced persons" or just "adults in Baidoa displacement camps"?

Response: We have revised the title to clarify our subject. The new title now reads "[Revised Title: Internally Displaced Persons in Baidoa Displacement Camps]."

2. Use a uniform font for the title.

Response: The title font has been standardized to ensure uniformity throughout the manuscript.

Introduction

1. Ensure a consistent referencing style throughout the manuscript.

Response: We have revised the manuscript to ensure consistent referencing style throughout.

2. In the sentence "Over the past few decades, the...," delete the bracket at the end of the statement.

Response: The bracket has been removed from the specified sentence.

3. Check the overall grammar of the text throughout the manuscript.

Response: We have conducted a thorough grammar check and corrected any errors found throughout the manuscript.

4. Regarding the burden of hypertension, provide more updated statistics on hypertension, using both global and regional data. Ensure a clear linkage and transition between the two because, as it stands right now, the statistics are scattered throughout the introduction, rendering it redundant.

Response: We have included updated global and regional statistics on hypertension and improved the linkage and transition between the two sets of data for better coherence.

5. Provide more context on the displaced populations and their specific vulnerabilities to hypertension to strengthen the rationale of the study. Discuss the factors therein.

Response: Additional context on the vulnerabilities of displaced populations to hypertension has been included, with a discussion of relevant factors contributing to this vulnerability.

6. The section would benefit from a discussion on the effects of hypertension.

Response: A new paragraph discussing the effects of hypertension has been added to the introduction.

7. Cite studies that have investigated hypertension among displaced populations, if any exist, or state the deficit if none.

Response: We have cited relevant studies investigating hypertension among displaced populations. Where studies are

lacking, we have noted this deficit to highlight the gap in the literature.

8. Discuss any interventions and strategies that have been implemented to tackle the problem of hypertension in these communities and state the possible gaps before your objective.

Response: A discussion on interventions and strategies addressing hypertension in displaced communities has been added, including a mention of the existing gaps.

Methods

1. Formatting issue: provide a heading for your Methods section.

Response: A clear heading for the Methods section has been added.

2. As stated above, there is a need to improve the overall grammar.

Response: We have reviewed and improved the grammar throughout the Methods section.

3. Provide more detail regarding the inclusion criteria. For instance, was there a specific displacement duration that was considered (ie, the minimum amount of time spent in the camp so far)?

Response: We have specified the inclusion criteria, including the consideration of the duration of displacement in the camps.

4. Provide a justification for the exclusion criteria.

Response: A justification for the exclusion criteria has been included to clarify the rationale behind them.

5. Provide the reference for "The sample size for this study was determined...."

Response: The appropriate reference for the sample size determination has been added to the Methods section.

6. Add more detail regarding the validation of the questionnaire. Was it adopted from previous studies? Was it pretested?

Response: We have provided additional details about the validation of the questionnaire, including its adoption from previous studies and the pretesting process.

7. Add detail on the measurement of blood pressure (BP). Who measured the BPs? Were they trained? How did you deal with white-coat hypertension? What was the interval between the different BP readings?

Response: We have added detailed information on blood pressure measurement, including who conducted the measurements, their training, how white-coat hypertension was addressed, and the intervals between readings.

Results

1. Again, appropriate headings should be provided. Check the grammar.

Response: Appropriate headings have been added to the Results section, and a grammar check has been completed.

2. Provide a more simplified and summarized Results section. For instance, "In this study, we enrolled 240 respondents, with a mean age...."

Response: The Results section has been simplified and summarized for clarity, including the specific example provided.

3. Table 1 is very confusing, especially the frequency and percentage columns. Clearly provide both the frequencies and percentages.

Response: Table 1 has been revised to clearly present both frequencies and percentages for better understanding.

4. Add a key for Figure 2 to give better representation or just integrate the data represented into the text.

Response: A key has been added to Figure 2 for clarity, and relevant data have been integrated into the text for additional context.

Discussion

1. Restate the objective at the start.

Response: The objective of the study has been restated at the beginning of the Discussion section.

2. Provide a concise summary of key findings.

Response: A concise summary of the key findings has been included at the beginning of the Discussion section.

3. Thoroughly discuss the implications of the factors found to be significantly associated with hypertension.

Response: A thorough discussion of the implications of the significant factors associated with hypertension has been added to the Discussion section.

References

1. Anonymous. Commentary on "Prevalence of Undiagnosed Hypertension Among Adult Displaced Individuals in Baidoa Camps, Somalia (Preprint)". JMIRx Med 2025;6:e71041. [doi: 10.2196/71041]
2. Jayte M. Prevalence of undiagnosed hypertension among adult displaced individuals in Baidoa camps, Somalia. medRxiv. Preprint posted online on Mar 26, 2024. [doi: 10.1101/2024.03.22.24304736]

Edited by E Meinert; submitted 18.12.24; this is a non–peer-reviewed article;accepted 18.12.24; published 03.06.25.

Please cite as:
Jayte M
Author's Response to a Commentary on "Prevalence of Undiagnosed Hypertension Among Adult Displaced Individuals in Baidoa Camps, Somalia (Preprint)"
JMIRx Med 2025;6:e70265
URL: https://xmed.jmir.org/2025/1/e70265
doi:10.2196/70265

Authors' Response to Peer Reviews of "Financial Feasibility of Developing Sustained-Release Incrementally Modified Drugs in Thailand's Pharmaceutical Industry: Mixed Methods Study"

Manthana Laichapis[1], PharmD; Rungpetch Sakulbumrungsil[1], PhD; Khunjira Udomaksorn[2], PhD; Nusaraporn Kessomboon[3], PhD; Osot Nerapusee[1], PhD; Charkkrit Hongthong[3], MSc; Sitanun Poonpolsub[4], PharmD

[1]Department of Social and Administrative Pharmacy, Faculty of Pharmaceutical Sciences, Chulalongkorn University, 254 Phayathai Road, Pathum Wan, Bangkok, Thailand

[2]Department of Social and Administrative Pharmacy, Faculty of Pharmaceutical Sciences, Prince of Songkla University, Songkla, Thailand

[3]Department of Social and Administrative Pharmacy, Faculty of Pharmaceutical Sciences, Khon Kaen University, Khon Kaen, Thailand

[4]Food and Drug Administration Thailand, Nonthaburi, Thailand

Corresponding Author:
Rungpetch Sakulbumrungsil, PhD
Department of Social and Administrative Pharmacy, Faculty of Pharmaceutical Sciences, Chulalongkorn University, 254 Phayathai Road, Pathum Wan, Bangkok, Thailand

Related Articles:

Companion article: https://www.medrxiv.org/content/10.1101/2024.07.29.24311184v1

Companion article: https://med.jmirx.org/2025/1/e78090

Companion article: https://med.jmirx.org/2025/1/e77627

Companion article: https://med.jmirx.org/2025/1/e65978

(JMIRx Med 2025;6:e77623) doi:10.2196/77623

KEYWORDS

financial; economics; R&D; research and development; surveys; interviews; costs; revenue; policies; drugs; pharmaceuticals

This is the authors' response to peer-review reports for "Financial Feasibility of Developing Sustained-Release Incrementally Modified Drugs in Thailand's Pharmaceutical Industry: Mixed Methods Feasibility Study."

Round 1 Review

Thank you for your valuable comments and for recognizing the importance of conducting incrementally modified drug (IMD studies. We appreciate your feedback and have made the necessary revisions to improve the clarity, depth, and quality of the paper [1]. Below are our responses to each point.

Reviewer H [2]

General Comments

This paper provides valuable insights into how the Thai pharmaceutical industry should prepare for future developments. The results can be used as a reference to support decision-making and to guide the definition of regulations and processes in Thailand.

Specific Comments

Major Comments

1. Methods: Could you elaborate on how the 5 incrementally modified drug (IMD) experts were selected? Additionally, why was the number of experts limited to 5?

Response: Thank you for your insightful question. We conducted in-depth interviews with 15 participants, ensuring data saturation in accordance with qualitative research methodology. Among them, 5 were local company owners specializing in IMD development, as they provided firsthand insights into industry challenges and opportunities. The remaining participants included experts from various sectors of IMD advancement, such as regulatory affairs, financial modeling, and clinical development, ensuring a comprehensive and diverse perspective. The selection criteria were designed to capture a balanced representation of stakeholders in the IMD landscape. Relevant details are provided in lines 101 - 102.

2. Tables 1 and 2: Please replace the term "Literature Review" with the specific author names and the corresponding year (Anno Domini).

Response: Thank you for your suggestion. We have replaced the term "literature review" with the specific author names and corresponding year where applicable. However, for sources derived from government documents and institutional reports, we have used the official abbreviations of the respective organizations to maintain clarity and accuracy.

3. Table 3: The values of US $1.46 million and US $18.6 million refer to the research and development costs only, correct? These values do not reflect the total cost of developing IMDs (refer to Table 2).

Response: Thank you for your inquiry regarding the values listed in Table 3. To clarify, the figures of US $1.46 million and US $18.6 million indeed represent comprehensive cost assessments. These values encompass the entirety of the research and development expenditures, which includes formulation development, clinical trials, production batches necessary for registration, and the registration process itself. The provided values are intended to reflect the total cost incurred up until the point of market authorization. We have ensured that these costs cover most, if not all, expenses associated with the development of IMDs before reaching market readiness. This clarification has been detailed in Table 2.

4. Since most of the numbers come from expert input, how do you ensure that these numbers are valid and accurately reflect real-world situations? It may be helpful to provide more information about the characteristics and qualifications of the key informants to support their credibility.

Response: Thank you for your thoughtful comment. To ensure the validity and real-world accuracy of expert-provided data, we applied a triangulation approach, incorporating insights from multiple sources, including literature reviews, surveys, and interviews. This cross-verification process enhanced the consistency and reliability of the findings. Additionally, the experts were selected based on their extensive experience and qualifications in drug development. They include industry leaders, policy makers, and researchers with direct involvement in IMD development and financial modeling. The relevant details can be found in lines 80 - 84 and 101 - 102. Please let us know if further clarification is needed.

Minor Comments

5. Please ensure that all abbreviations are defined the first time they appear in the document. For example, "IMD" should be written out as "Innovative Medical Devices (IMD)" when it is first mentioned, particularly in the introduction.

Response: Thank you for your feedback. We have reviewed the document and ensured that all abbreviations are properly defined upon first mention.

Reviewer BK [3]

General Comments

This paper presents a thorough analysis of the financial feasibility of developing incrementally modified drugs (IMDs) within the Thai pharmaceutical industry. It aligns well with Thailand's National Strategic Master Plan and provides valuable insights for stakeholders regarding investment decisions and policy development. The mixed- methods approach, including financial modeling, surveys, and interviews, lends credibility to the findings, while the focus on sustained-release dosage forms highlights a specific and practical application. The paper is well- structured and contributes meaningfully to the discussion on enhancing local pharmaceutical capabilities. However, there are areas where clarity, presentation, and depth can be improved to strengthen its impact.

Specific Comments

Major Comments

1. Clarity in objectives: While the paper provides an extensive background on Thailand's pharmaceutical landscape, the research objectives could be more explicitly stated at the beginning of the introduction to guide the reader more effectively.

Response: Thank you for your suggestion to enhance the clarity of the research objectives. We have revised the introduction to clearly and explicitly state the research objectives at the beginning, providing better guidance for the reader and improving the overall clarity of the study's purpose.

2. Discussion of results: The discussion section could delve deeper into comparing the financial feasibility of IMDs with other pharmaceutical products, especially generic drugs, to highlight the broader implications of the findings.

Response: Thank you for your valuable suggestion on comparing IMDs with other pharmaceutical products. We have expanded the discussion section to provide a more in-depth comparison of the financial feasibility of IMDs with new drugs, new generic drugs, and the US Food and Drug Administration 505(b)(2) New Drug Application program, enhancing the applicability of the findings. The revisions can be found in lines 191 - 199.

3. Policy recommendations: Although the paper suggests policy recommendations, it would benefit from providing concrete examples of how these policies have been successfully implemented in other regions or industries. This would add depth and context to the recommendations.

Response: Thank you for your valuable feedback on the policy recommendations section of our manuscript. We acknowledge your suggestion to enhance this section by providing concrete examples of successful policy implementations from other regions or industries. However, given the primary focus of our study on the financial aspects of developing IMDs within Thailand's pharmaceutical industry, we have revised the manuscript to refine the scope of our conclusions. In this revision, we have removed detailed policy recommendations. Instead, we now suggest that the findings could be beneficial for planning strategic support within the industry. This adjustment helps to maintain the focus on the financial analysis and ensures that the recommendations are directly supported by our research findings without extending beyond the evidence provided. We believe this approach will keep the study concise and focused on its primary objectives.

4. References and citation quality: The paper relies on only 15 references, which is insufficient for a study of this scope. Furthermore, only a few of these references are from peer-reviewed scientific journals, while the rest are reports and secondary sources. This significantly weakens the academic foundation of the study. It is strongly recommended to update the references section by incorporating recent, high-quality, and peer-reviewed articles.

Response: Thank you for highlighting this weakness in our study. We have strengthened its academic foundation by incorporating additional high-quality, peer-reviewed articles. However, as IMD remain a relatively new topic with limited peer-reviewed literature available, we primarily relied on in-depth interviews as the main methodology for estimating costs and key parameters.

Minor Comments

5. Terminology consistency: Terms like "incrementally modified drugs" and "IMDs" should be consistently used throughout the text to avoid confusion.

Response: Thank you for your feedback. We have reviewed the document and ensured that all abbreviations are properly defined upon first mention.

6. Figures and tables: Ensure all figures and tables are adequately labeled and referenced in the text. For instance, the presentation of financial data could be enhanced with clearer visualizations.

Response: Thank you for your valuable suggestion. We have revised all three tables for improved clarity and ensured that they are properly referenced throughout the text.

7. Formatting and grammar: Minor grammatical errors and formatting inconsistencies (eg, use of citations and spacing) should be addressed for a polished presentation.

Response: Thank you for highlighting this point. We have carefully reviewed the document to correct formatting inconsistencies, improve citation accuracy, and ensure grammatical correctness.

8. Abstract refinement: The abstract could be more concise, emphasizing key findings and policy implications without overly detailed descriptions of methods.

Response: Thank you for your feedback. We have revised the abstract into a structured format, making it more concise while emphasizing key findings.

9. Future research directions: Including a section on future research directions would enhance the paper's utility for academics and policy makers.

Response: Thank you for your valuable feedback on future research directions. As we mentioned earlier, IMDs are relatively new, presenting numerous research opportunities. In response, we have added a future research directions section, offering insights into the development of IMDs from patient, regulatory, and market-access perspectives. This addition provides valuable data for policy makers and the industry. The revisions are reflected in lines 216 - 223.

We appreciate the detailed feedback, which has significantly improved the clarity, structure, and academic rigor of our study. Please let us know if further refinements are needed.

References

1. Laichapis M, Sakulbumrungsil R, Udomaksorn K, et al. Financial feasibility of developing sustained-release incrementally modified drugs in Thailand's pharmaceutical industry: mixed methods feasibility study. JMIRx Med 2025;6:e65978. [doi: 10.2196/65978]
2. Luksameesate P. Peer review of "Financial Feasibility of Developing Sustained-Release Incrementally Modified Drugs in Thailand's Pharmaceutical Industry: Mixed Methods Feasibility Study". JMIRx Med 2025;6:e78090. [doi: 10.2196/78090]
3. Shkarupeta E. Peer review of "Financial Feasibility of Developing Sustained-Release Incrementally Modified Drugs in Thailand's Pharmaceutical Industry: Mixed Methods Feasibility Study". JMIRx Med 2025;6:e77627. [doi: 10.2196/77627]

Abbreviations

IMD: incrementally modified drug

Edited by A Grover; submitted 16.05.25; this is a non–peer-reviewed article;accepted 16.05.25; published 01.07.25.

Please cite as:
Laichapis M, Sakulbumrungsil R, Udomaksorn K, Kessomboon N, Nerapusee O, Hongthong C, Poonpolsub S
Authors' Response to Peer Reviews of "Financial Feasibility of Developing Sustained-Release Incrementally Modified Drugs in Thailand's Pharmaceutical Industry: Mixed Methods Study"
JMIRx Med 2025;6:e77623
URL: https://xmed.jmir.org/2025/1/e77623
doi:10.2196/77623

Authors' Response to Peer Reviews of "Improving Tuberculosis Detection in Chest X-Ray Images Through Transfer Learning and Deep Learning: Comparative Study of Convolutional Neural Network Architectures"

Alex Mirugwe[1], BSc, MSci; Lillian Tamale[2], PhD; Juwa Nyirenda[3], PhD

[1]School of Public Health, Makerere University, Kawalya Kaggwa Close, Plot 20A, Kampala, Uganda

[2]Faculty of Science and Technology, Victoria University, Kampala, Uganda

[3]Department of Statistical Sciences, University of Cape Town, Cape Town, South Africa

Corresponding Author:
Alex Mirugwe, BSc, MSci
School of Public Health, Makerere University, Kawalya Kaggwa Close, Plot 20A, Kampala, Uganda

Related Articles:

Companion article: https://www.medrxiv.org/content/10.1101/2024.08.02.24311396v1

Companion article: https://med.jmirx.org/2025/1/e 77171

Companion article: https://med.jmirx.org/2025/1/e 77174

Companion article: https://med.jmirx.org/2025/1/e66029

(JMIRx Med 2025;6:e77221) doi:10.2196/77221

KEYWORDS

tuberculosis detection; tuberculosis; TB; chest x-ray classification; diagnostic imaging; radiology; medical imaging; convolutional neural networks; data augmentation; deep learning; early warning; early detection; comparative study

This is the authors' response to peer-review reports for "Improving Tuberculosis Detection in Chest X-Ray Images Through Transfer Learning and Deep Learning: Comparative Study of Convolutional Neural Network Architectures."

Round 1 Review

Reviewer AE [1]

General Comments

Clarity and Structure

The paper [2] presents a comprehensive overview of the methods and results but can benefit from clearer transitions between sections. For instance, adding brief connecting sentences at the end of each section would help guide the reader into the next topic.

Consider reorganizing the "Discussion" section to first summarize the key findings before delving into their implications. This will reinforce the reader's understanding of the main outcomes.

Writing Style

Aim for more active voice usage to enhance readability. For example, change "It was observed that VGG16 outperformed other models" to "We observed that VGG16 outperformed other models."

Simplify overly technical or long sentences to improve readability. Breaking complex sentences into two simpler ones can make the content easier to follow.

Response: We have revised the manuscript to improve transitions between sections by adding concluding statements that summarize key points and guide the reader to the next section. Regarding the Discussion section, we believe the current structure effectively presents the findings and their implications. The key outcomes are already summarized at the start of the section, followed by a detailed discussion of their clinical and technical implications.

Specific Comments by Section

Abstract

Sentence clarification: The phrase "necessitating more efficient and accurate diagnostic methods" could be expanded to briefly indicate why current methods are insufficient.

Results detail: When mentioning model performance, briefly state why VGG16's superior performance is significant compared to others.

Response: We have revised the abstract to enhance its clarity and readability. Additionally, we included a clear Objective section to directly address the comment and make the study's purpose more explicit.

For sentence clarification, we have revised the Introduction section of the abstract to clearly indicate why current diagnostic methods are insufficient. For results details, we have revised the Results section of the abstract to explain why VGG16's superior performance was significant, emphasizing its balance of diagnostic accuracy and computational efficiency.

Introduction

Background information: The explanation of the global tuberculosis (TB) burden is informative, but it could benefit from briefly mentioning current limitations in artificial intelligence–based TB detection in developing countries. Motivation clarification: Ensure that the motivation for choosing specific convolutional neural network architectures is clearly linked to gaps in existing literature.

Response: We have revised the Introduction section to expand on the paragraphs, addressing the limitations of artificial intelligence–based TB detection in developing countries and clarifying the motivation for choosing specific convolutional neural network (CNN) architectures.

Methods

Preprocessing details: The detailed explanation of normalization and data augmentation is excellent, but it might be beneficial to briefly mention how these choices align with previous research findings or unique aspects of this study.

Transfer learning: Include a brief comparison of why transfer learning was chosen over training models from scratch.

Response: We have revised the Pre-Processing section to incorporate findings from previous research in the Normalization and Data Augmentation subsections, emphasizing how these techniques address unique aspects of this study, such as dataset imbalance and real-world variability in chest x-ray data. For the Transfer Learning section, we added a brief comparison explaining why transfer learning was preferred over training models from scratch, highlighting its advantages in resource-limited settings and its proven effectiveness in medical imaging tasks.

Results

Visualization: The table summarizing model performance is comprehensive, but consider including a concise narrative to describe key trends observed in the data.

Analysis clarification: When discussing why data augmentation did not enhance performance, elaborate on how this aligns with or contradicts findings from other studies.

Discussion

Comparison with previous studies: Add a few sentences comparing the results with existing studies that used the same models or datasets to provide context.

Implications: Discuss the practical implications of using VGG16 in resource-constrained environments where computational efficiency is crucial.

Conclusion

Highlight novelty: Emphasize what makes this study's approach unique, such as the use of specific architectures on a larger dataset, and how this adds to the current body of knowledge.

Future work suggestions: Include more detailed recommendations for future studies, potentially suggesting how to further leverage data augmentation strategies.

Response: We have revised the Discussion section to include two additional paragraphs elaborating on why data augmentation did not improve performance. These paragraphs provide a detailed explanation of how our findings align with certain previous studies while contrasting with others.

Reviewer AI [3]

1. The dataset includes a large imbalance between TB-positive and TB-negative images (700 vs 3500). Explain how this imbalance was addressed beyond augmentation or whether balancing techniques like oversampling were considered.

Response: No additional balancing methods were used, such as oversampling or undersampling. Instead, data augmentation was specifically used to introduce variability and enhance the representation of TB-positive images, constituting the smaller class. Given the study's objectives and dataset characteristics, this approach was considered adequate for addressing the class imbalance.

2. While each architecture's parameters are listed, there is no in-depth discussion on why these specific parameters (eg, dropout rates, learning rates) were selected.

Response: A paragraph has been added at the end of the CNN Architectures subsection to explain how we arrived at the parameters used for training. This addition clarifies that the parameters were determined through a rigorous iterative process of experimentation and were selected based on their ability to deliver optimal performance across the evaluated architectures.

3. The conclusion that data augmentation did not improve performance lacks specific references to possible reasons.

Response: We have added a detailed explanation in the Discussion section, citing studies that achieved similar results and those with augmentation improved performance. We have also explained why the latter was not the case in our study.

4. While computational time for each model is reported, further analysis of the practical implications, such as cost-effectiveness for clinical settings, is missing.

Response: In response to the comment regarding the practical implications of computational time, we have added a paragraph in the Discussion section to address cost-effectiveness and the relevance of model training times for clinical settings.

5. The manuscript mentions transfer learning with pretrained ImageNet weights, but there is limited information on why this was the chosen approach versus training from scratch.

Response: We added a brief comparison explaining why transfer learning was preferred over training models from scratch, highlighting its advantages in resource-limited settings and its proven effectiveness in medical imaging tasks.

6. Throughout the Results section, adding comparative charts or visual aids for each model's performance across metrics like accuracy, precision, and area under the receiver operating characteristic curve would improve readability.

7. The Conclusion could benefit from a clearer statement on how these findings advance the field of TB detection in medical imaging.

Response: Your suggestions have been addressed by adding more clarity to the Results, Discussion, and Conclusion sections.

References

1. Pitakaso R. Peer review of "Improving Tuberculosis Detection in Chest X-Ray Images Through Transfer Learning and Deep Learning: Comparative Study of Convolutional Neural Network Architectures". JMIRx Med 2025;6:e77171. [doi: 10.2196/77171]
2. Mirugwe A, Tamale L, Nyirenda J. Improving Tuberculosis Detection in Chest X-Ray Images Through Transfer Learning and Deep Learning: Comparative Study of Convolutional Neural Network Architectures. JMIRx Med 2025;6:e66029. [doi: 10.2196/66029]
3. Nanthasamroeng N. Peer review of "Improving Tuberculosis Detection in Chest X-Ray Images Through Transfer Learning and Deep Learning: Comparative Study of Convolutional Neural Network Architectures". JMIRx Med 2025;6:e77174. [doi: 10.2196/77174]

Abbreviations

CNN: convolutional neural network
TB: tuberculosis

Edited by S Amal; submitted 09.05.25; this is a non–peer-reviewed article;accepted 09.05.25; published 01.07.25.

Please cite as:
Mirugwe A, Tamale L, Nyirenda J
Authors' Response to Peer Reviews of "Improving Tuberculosis Detection in Chest X-Ray Images Through Transfer Learning and Deep Learning: Comparative Study of Convolutional Neural Network Architectures"
JMIRx Med 2025;6:e77221
URL: https://xmed.jmir.org/2025/1/e77221
doi:10.2196/77221

Authors' Response to Peer Review of "Using Electrooculography and Electrodermal Activity During a Cold Pressor Test to Identify Physiological Biomarkers of State Anxiety: Feature-Based Algorithm Development and Validation Study"

Jadelynn Dao[1*], BSc; Ruixiao Liu[2], BSc; Sarah Solomon[3], BSc, MD; Samuel Aaron Solomon[2*], BSc, MEng, PhD

[1]Computer Science, California Institute of Technology, Pasadena, CA, United States

[2]Medical Engineering, California Institute of Technology, 1200 E California Blvd, Pasadena, CA, United States

[3]Adult Psychiatry, Dartmouth College, Hanover, NH, United States

[*]these authors contributed equally

Corresponding Author:
Samuel Aaron Solomon, BSc, MEng, PhD
Medical Engineering, California Institute of Technology, 1200 E California Blvd, Pasadena, CA, United States

Related Articles:

Companion article: https://arxiv.org/abs/2411.17935v1

Companion article: https://med.jmirx.org/2025/1/e72093

Companion article: https://med.jmirx.org/2025/1/e69472

(JMIRx Med 2025;6:e77440) doi:10.2196/77440

KEYWORDS

stress; biomarker discovery; EOG; EEG; medical informatics; electrooculography; electroencephalogram

This is the authors' response to the peer-review report of "Using Electrooculography and Electrodermal Activity During a Cold Pressor Test to Identify Physiological Biomarkers of State Anxiety: Feature-Based Algorithm Development and Validation Study."

List of Major Concerns and Feedback

Concerns With Methods

It would be helpful to document the name of the device and manufacturer used in this study to record the electrooculography (EOG). This would help other researchers who may want to reproduce the results.

Response: We appreciate the reviewer's [1] suggestion and agree that providing this information would improve the reproducibility and clarity of our study [2]. We have now added the name of the EOG device and its manufacturer in the Methods section of the revised manuscript. The updated text reads as follows:

- "Our setup integrated the AD8232 (Analog Devices), a biopotential amplifier designed to capture physiological signals, which we optimized for measuring EOG activity."
- "Additionally, 19 trials lasting between 30 seconds and 2 minutes were conducted under conditions with no blinking, but with deliberate wire movements introduced by manually adjusting or lightly tugging the electrode leads."
- "EOG recording used the same setup as the Blink Identification EOG Dataset (BLINKEO) data collection. Electrodes were positioned above and below one eye to detect vertical eye movements by capturing corneo-retinal potential shifts."

Similarly, it would be helpful to add additional details about the cold pressor test (CPT) methods. For example, was a commercially available circulating water bath used to maintain a constant water temperature? Was the temperature of the subject's hand monitored? The details of the cold stressor test (the water temperature, the period of immersion, and the cutoff point) should be added for the sake of clarity, transparency, and reproducibility. Past studies using these metrics should also be referenced for details (eg, [3]). These methodological details may also be added in the form of a figure to add clarity to the experimental setup.

Response: In response, we have expanded the Methods section to include additional details about the CPT setup. First, the reference that was suggested in the reviewer's comment was added. In response, we have also expanded the Methods section to provide a clearer description of the CPT protocol. Specifically, we now included "In the cold-water trials, participants immersed their hand in a circulating water bath set

to a constant temperature of 0-6°C. Participants maintained immersion for approximately 5 minutes or until voluntary withdrawal." Furthermore, we have removed mention of exercise trials, as they were not used in dataset creation or analysis and are thus not relevant to the study.

To better understand the individual response to the cold challenge before participating in the actual experiment, it is advised that the manuscript states what type of participant testing was or was not adopted in the cold pressor testing experiment. For example, what were the tolerance times? Were there any gender differences? If any pretesting data were collected, analyzing them and presenting them as results would add clarity to the results.

Response: We did not implement a formal pretesting phase to assess individual tolerance times before the experiment. All participants were instructed to immerse their hand in the CPT until they reached their tolerance limit or approximately 5 minutes (300 seconds). A summary of trial durations for each phase of the experiment—baseline (before hand submersion), CPT (cold water immersion), and recovery (after hand removal)—is presented in Table 2c. This table includes the minimum, 25th percentile, median, 75th percentile, and maximum tolerance times recorded across participants. Table 2c's description was amended to make this more clear:

- "d. Summary of the duration of time EDA and EOG features are collected from, across different experimental phases. For each phase—Baseline (before hand submersion), Cold Pressor Test (cold water immersion), and Recovery (after hand removal) —both tables list the minimum, 25th percentile, median, 75th percentile, and maximum duration (in seconds)."

Regarding gender differences, our study was not explicitly designed to analyze gender-based variations in cold stress tolerance, and the sample size for gender-based comparisons is limited. However, we acknowledge the potential relevance of such analyses and have noted this as an area for future investigation in the Conclusion:

- "An important next step is to investigate potential gender-based and race-based differences in physiological responses to acute stress and our current methods of inducing stress, as our current study was not explicitly designed for such analysis but acknowledges its relevance."

It is unclear if the 65 repeating blinking trials and the 19 no-blinking trials were collected from the same individual or from different individuals. Please clarify.

Response: We agree that clarifying whether the trials were conducted on the same or different individuals improves the transparency of our methodology. In the revised manuscript, we have explicitly stated that all trials were conducted on the same individual to ensure consistency in signal characteristics. The updated text now reads "All trials were conducted on the same two individuals for consistency in signal characteristics."

No signal voltage/electrical records for electrodermal activity [EDA] were found in the manuscript. Is this intentional? Please consider adding this information.

Response: In the revised manuscript, we have now explicitly provided details on the EDA signal acquisition, including the applied voltage and electrical characteristics. The updated text reads as follows:

- "EDA signals were recorded using a GSR (Galvanic Skin Response) sensor with MCP606 (Microchip Technology) operational amplifiers, operating at an excitation voltage of 0.5V to measure skin conductance. Electrodes were placed on the forehead, chosen for its sensitivity to stress-induced sweat gland activity. The recorded signals were digitized and processed in real-time using an ESP32-S3 WROOM-1 (Espressif Systems) microcontroller, which managed data acquisition, signal processing, and wireless transmission."

It would be important to add details of ordinal variables present in the Positive and Negative Affect Schedule (PANAS) and the State-Trait Anxiety Inventory (STAI-State), and clearly state their function and use in Supplementary Table 2.

Response: In response, we have updated Supplementary 2's table to explicitly describe how these scales function in the assessment of emotional and anxiety states. The revised descriptive text in Supplementary 2 now reads:

- "The survey items from the Positive and Negative Affect Schedule (PANAS) and the State-Trait Anxiety Inventory (STAI-State) were used to assess participants' emotional and anxiety responses during the experiment. The PANAS scale consists of 10 items measuring Positive Affectivity and Negative Affectivity, each rated on a 1-5 Likert scale, where higher scores indicate stronger affective states. The STAI-State consists of 20 items assessing state anxiety, measured on a 1-4 Likert scale, where responses indicate varying degrees of agreement with statements reflecting anxiety levels. Higher scores in negative affectivity and anxiety-related items indicate greater distress, while higher scores in positive affectivity items indicate greater emotional well-being."

Concerns With Analysis

F_1-scores that were mentioned in the text (87.34% and 79.99%) are not present within the figures. Moreover, an F_1-score is an integer value from 0 to 1, taking precision and recall into account, and is not often expressed as a percentage.

Response: The updated text now expresses the F_1-scores as decimal values, aligning with the conventional representation. In addition, the figures now include the accuracy and F_1-score: "0.8734" and "0.7999."

Figure 1c has two separate graphs; it should be captioned as 1c and 1d. What do both these graphs portray? The second graph for 1c is missing titles for the x- and y-axes—the current assumption is that they are the same as the first graph.

Response: The figure has been updated to distinctly label the two separate graphs as Figure 1c and Figure 1d in both the figure and the caption. We clarified the purpose of both graphs, stating that they each depict independent blink events, highlighting the variability in peak shape that can occur in EOG recordings:

- "d. Another example of a blink peak, demonstrating the variability in blink peak shapes observed across recordings. The feature extraction process remains consistent, with boundaries determined by identifying the nearest minima on either side of the peak."

Table 1 lacks a legend and is shown as panel a of Table 2. Please check how the tables are referenced in the text to make sure they reference the right one.

Response: Table 1 is now correctly referenced in the manuscript to ensure clarity. A brief description has been included to clarify its contents, explicitly stating that it summarizes the trial characteristics, total duration, and peak detection results before and after filtering.

- "Table 1 summarizes the characteristics of these trials, including session count, total recording time, and peak detection results before and after filtering."

We have verified all text references to ensure that Table 1 and Table 2 are cited appropriately.

- "Sixteen participants (N=16) between ages 26-31 took part in the study, and demographic information, including race and gender, was collected and is summarized in Table 2a-b. Each trial lasted about 10-15 minutes and was divided into three phases: baseline, CPT (Cold Pressor Test), and recovery. The length of the trial and the data used for feature analysis is as detailed in Table 2c-d."

The captions of the figures should have statistical information when relevant. For example, in Figure 3, the caption should include a description of what data were plotted and the meaning of the graph. Presumably plotting medians, quartiles, and SDs? Also, please report n values.

Response: Figure 3 has been updated to include the median and SD of each score. Figure 2 has been updated to include accuracy and F_1-score for each culling step.

Concerns With Ethics

It is not clear what the ethical statement at the end of the manuscript, which states that the study was exempt from review board approval, means. That statement should be revised for clarification. In addition, details regarding whether or not institutional review board approval was obtained, whether the study involved consenting participants and used humans, how the data were collected and used, how the data were handled to protect the privacy of study participants, and any other ethical procedures that were followed to protect subjects from any harm due to participation in the study should be added.

Response: We have clarified the ethical statement at the end of the manuscript. This study was conducted in accordance with ethical guidelines for research involving human participants. All patient data were fully anonymized prior to analysis, with identifying information removed and data transmission secured using byte-splicing encryption methods. All participants provided informed consent for the use of their data in this study. The study adhered to data privacy and security protocols to ensure the confidentiality and protection of participants.

List of Minor Concerns and Feedback

Minor Concerns With Methods

Please document whether the data were taken from each subject only once or whether data were obtained several times from a subject.

Response: In response, we have explicitly stated that data were collected from each subject only once in the revised manuscript. The updated text now reads:

- "Sixteen participants (N=16) between ages 26-31 took part in the study, and demographic information, including race and gender, was collected and is summarized in Table 2a-b. Data was taken from each subject only once."

Referring to the line "To focus on blink-like events, we applied criteria based on established blink characteristics," the criteria used to establish blink characteristics should be cited, if not already given.

Response: To address this, we have now clarified how we derived this criteria. The revised text now references the methodology of BLINKER, a pipeline for extracting ocular indices such as blink rate, blink duration, and blink velocity-amplitude ratios from electroencephalogram channels, EOG channels, and/or independent components.

Shapley additive explanations (SHAP) analysis was performed on combinations of 5 features. Please clarify on what basis these 5 features were chosen (out of 15 of EDG and 33 of EOG).

Response: We have clarified the description of the SHAP analysis methodology:

- "In this study, SHAP analysis was performed on combinations of five features, selected from the total feature set of 15 EDG and 33 EOG features, highlighting the significance of how certain biomarkers, used together, reveal more prominent interactions and effects on model predictions. This approach underscores that certain biomarkers, while potentially less impactful individually, can demonstrate substantial importance when analyzed as part of a group. By evaluating these interactions, we understand how combinations of features can provide insights into the model's behavior that single-feature analyses might overlook."
- "The quality of a set of features is determined by considering their collective contribution to the model's predictions, measured through the mean absolute SHAP values across the dataset. A high-quality set of features is one where the combination of features demonstrates substantial importance, as indicated by a higher mean absolute SHAP values. This benchmark reflects not only the magnitude of individual contributions but also the degree to which the features, as a group, interact to enhance the predictive power of the model."

Minor Concerns With Analysis and Presentation

Page 10, Electrooculography (EOG) Signal Segmentation section: the authors mentioned that they extracted 33 features;

however, Supplementary 4 mentioned 35 feature definitions. Please revise and correct.

Response: We have cross-checked the manuscript and Supplementary 4 to ensure consistency in the reported number of features. A total of 35 features were used, so we have revised the EOG Signal Segmentation section to correctly state "35 features" instead of "33."

In Figure 3, please put "STAI-State survey score" on the y-axis for clarification rather than just "Scores." In addition to box and whiskers plots, adding column graphs for positive affectivity, negative affectivity, and s-anxiety might be beneficial to more clearly express the SD present within the data.

Response: We agree that column graphs can effectively complement the box plots by visually emphasizing SDs within the dataset. We have introduced bar charts with error bars to represent mean survey scores for each stage (baseline, CPT, and recovery). The axes and labels were also clarified, per request. The figure description now includes:

- "Figure 3 User-reported survey responses during each stage of the trial, displaying both box-and-whisker plots and column graphs for Positive Affectivity, Negative Affectivity, and State Anxiety (S-Anxiety) across the Baseline, CPT, and Recovery stages."

It would be beneficial to graphically display the F_1-scores that were collected across the study.

Response: We have updated Figure 2 to include the F_1-scores across each step of the culling pipeline.

The figures are quite small, which makes readability a little difficult. Please make the text larger to improve readability and accessibility.

Response: Figure axes labels, headings, and some descriptions were adjusted with larger text.

The Figure 1a description states, "The red dotted lines indicate the center of the peak…," but these appear to be gray.

Response: We have resolved this figure description, which now reads, "The grey dotted lines…"

Suggestions

Consider the inclusion of a Limitations section in this manuscript to better discuss potential limitations due to the skewness in male and female participants, data curation, applied methodologies, and other limitations of the study.

Response: A "Limitations" section was added to this manuscript in the Conclusion. It reads "This study advances state anxiety biomarker detection using Electrooculography (EOG) and Electrodermal Activity (EDA), but several limitations should be noted. The participant pool (N=16) was demographically skewed, with a predominance of male and Asian participants, limiting generalizability. Data was collected only once per subject, preventing analysis of intra-individual variability over time. Future studies should incorporate larger and more diverse populations with longitudinal data. "The Cold Pressor Test (CPT) was conducted in a controlled lab environment, which may not fully reflect real-world anxiety triggers. Additionally, motion artifacts in EOG recordings, despite filtering efforts, could impact signal clarity. EDA signals were recorded using a single forehead electrode, though different placements (e.g., fingertips) may improve accuracy. Improved artifact detection and additional motion-tracking sensors could enhance data quality. Feature selection for SHAP analysis focused on optimizing interpretability, but alternative selections may yield different insights. Models and analyses constructed using this dataset may not generalize well to other stress-inducing scenarios. External validation using independent datasets is necessary to confirm these findings."

A figure showing the trial structure would be very useful to understand how the data were collected.

Response: The design of these trials facilitated the collection of time series data during an environmental stressor. We have added an additional figure to make the setup/timeline of this experiment more clear:

- "Figure 2 This figure presents a visual representation of the experiment timeline, detailing the Baseline, Cold Pressor Test (CPT), and Recovery phases. The raw Electrooculography (EOG) and Electrodermal Activity (EDA) signals across these phases show no immediately clear trend distinguishing the baseline and recovery from the CPT stressor. However, when specific features such as Blink Duration from EOG and Hjorth Activity from EDA are extracted and overlaid, more distinct patterns emerge, and can be used to quantify physiological responses to stress induction and subsequent recovery."

References

In the third paragraph of the Introduction, adding a reference to other techniques used to provoke anxiety, including the reduced EDA response in depressed patients, and the conflicting studies could be helpful to the readers.

Response: Four additional references were made to cite techniques that have been shown to provoke anxiety. Also, additional sentences were added to discuss the response variability introduced by depression, medication usage, and methodological differences:

- "Electrodermal activity (EDA) is a common measure of physiological arousal, but its reliability in depression research remains debated. Some studies report reduced EDA responses in individuals with major depressive disorder, suggesting impaired autonomic reactivity12 and emotional hypo-responsiveness13. However, conflicting findings point to variability due to factors like medication use and methodological differences14, emphasizing the need for further research on the relationship between physiological signals and emotional states."

In the Introduction, fourth paragraph, the reference "Schachter and Singer" is not present in the References. Is this the wrong reference, or it just needs to be added to the list?

Response: Schachter and Singer [4] has now been added to the list of references.

In the Introduction, third page, third paragraph, it is advised to add references to document the reduced EDA response in depressed patients and the conflicting studies.

Response: This comment is a repeat of the first comment in the References section and was addressed accordingly.

In the Methods, please cite sources for the Butterworth filter (page 5), the Savitzky-Golay filter (page 5), and all other analyses.

Response: Specifically, we now reference Virtanen et al [5] for the implementation of these filters in the SciPy library. Additional citations have been included where applicable to provide proper attribution for the analytical techniques used.

Reference 2: Include full citation with a link.

Response: This was corrected.

Reference 3: It is advised to correct the article name to "APA 2023 Stress in America Topline Data."

Response: This was corrected.

Reference 4: The correct citation should be "Kazanskiy NL., Khonina S.N., Butt M.A. A review on flexible wearables—Recent developments in non-invasive continuous health monitoring. Sens. Actuators A Phys. 2024;366:114993. doi: 10.1016/j.sna.2023.114993."

Response: This was corrected.

Reference 10: The correct citation should be: "Electrooculogram Analysis and Development of a System for Defining Stages of Drowsiness Master's Thesis Project in Biomedical Engineering, Linköping University, Dept. Biomedical Engineering, LiU-IMT-EX-351 Linköping 2003. Available: https://www.diva.portal.org/smash/get/diva2:673960/FULLTEXT01.pdfTest."

Response: This is now reference 16. This was corrected.

Reference 19: The correct citation should be "Anxiety Detection Using Multimodal Physiological Sensing, 2021 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI), Athens, Greece, 2021, pp. 1-4, doi: 10.1109/BHI50953.2021.9508589."

Response: This is now reference 25. This was corrected.

Reference 23: Revising this citation is advised as searching on the internet shows error 404. The requested URL was not found on this server. Moreover, this is not a proper citation—give the edition number of the book (there are at least 5 editions) and publication year, as well as the page number of the cited data point about typical blink elapsed time.

Response: This is now reference 29. This was corrected.

Reference 27: The correct citation should be "Hassanein, A.M.D.E., Mohamed, A.G.M.A. & Abdullah, M.A.H.M. Classifying blinking and winking EOG signals using statistical analysis and LSTM algorithm. Journal of Electrical Systems and Inf Technol 10, 44 (2023). https://doi.org/10.1186/s43067-023-00112-2"

References

1. Saderi D, Rasania S, Olatoye T, et al. Peer review of "State Anxiety Biomarker Discovery: Electrooculography and Electrodermal Activity in Stress Monitoring (Preprint)". JMIRx Med 2025;6:e72093. [doi: 10.2196/72093]
2. Dao J, Liu R, Solomon S, Solomon SA. Using Electrooculography and Electrodermal Activity During a Cold Pressor Test to Identify Physiological Biomarkers of State Anxiety: Feature-Based Algorithm Development and Validation Study. JMIRx Med 2025;6:e69472. [doi: 10.2196/69472]
3. Mitchell LA, MacDonald RAR, Brodie EE. Temperature and the cold pressor test. J Pain 2004 May;5(4):233-237. [doi: 10.1016/j.jpain.2004.03.004] [Medline: 15162346]
4. Schachter S, Singer JE. Cognitive, social, and physiological determinants of emotional state. Psychol Rev 1962 Sep;69:379-399. [doi: 10.1037/h0046234] [Medline: 14497895]
5. Virtanen P, Gommers R, Oliphant TE, et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat Methods 2020 Mar;17(3):261-272. [doi: 10.1038/s41592-019-0686-2] [Medline: 32015543]

Abbreviations

BHI: Biomedical and Health Informatics
BLINKEO: Blink Identification EOG Dataset
CPT: cold pressor test
EDA: electrodermal activity
EOG: electrooculography
GSR: galvanic skin response
PANAS: Positive and Negative Affect Schedule
SHAP: Shapley additive explanations
STAI-State: State-Trait Anxiety Inventory

Edited by A Schwartz; submitted 13.05.25; this is a non–peer-reviewed article;accepted 12.05.25; published 10.07.25.

Please cite as:
Dao J, Liu R, Solomon S, Solomon SA
Authors' Response to Peer Review of "Using Electrooculography and Electrodermal Activity During a Cold Pressor Test to Identify Physiological Biomarkers of State Anxiety: Feature-Based Algorithm Development and Validation Study"
JMIRx Med 2025;6:e77440
URL: https://xmed.jmir.org/2025/1/e77440
doi:10.2196/77440

Authors' Response to Peer Reviews of "Advancing Early Detection of Major Depressive Disorder Using Multisite Functional Magnetic Resonance Imaging Data: Comparative Analysis of AI Models"

Masab Mansoor[1], DBA; Kashif Ansari[2], MD

[1]School of Medicine, Edward Via College of Osteopathic Medicine, Louisiana Campus, 4408 Bon Aire Dr, Monroe, LA, United States

[2]East Houston Medical Center, Houston, TX, United States

Corresponding Author:
Masab Mansoor, DBA
School of Medicine, Edward Via College of Osteopathic Medicine, Louisiana Campus, 4408 Bon Aire Dr, Monroe, LA, United States

Related Articles:

Companion article: https://www.medrxiv.org/content/10.1101/2024.08.13.24311933v1

Companion article: https://med.jmirx.org/2025/1/e76744

Companion article: https://med.jmirx.org/2025/1/e76746

Companion article: https://med.jmirx.org/2025/1/e76747

Companion article: https://med.jmirx.org/2025/1/e65417

(JMIRx Med 2025;6:e75617) doi:10.2196/75617

KEYWORDS

major depressive disorder; machine learning; functional MRI; early detection; artificial intelligence; psychiatry

This is the authors' response to peer-review reports of "Advancing Early Detection of Major Depressive Disorder Using Multisite Functional Magnetic Resonance Imaging Data: Comparative Analysis of AI Models."

Round 1 Review

We thank the editors and reviewers for their thoughtful and constructive feedback on our manuscript "Advancing Early Detection of Major Depressive Disorder Using Multi-site Functional Magnetic Resonance Imaging Data: Comparative Analysis of AI Models" [1]. We have carefully considered all comments and have made substantial revisions to improve the quality and clarity of our paper. Below, we address each point raised by the reviewers.

Anonymous [2]

Major Comments

Interpretability of artificial intelligence (AI) models: While the paper discusses the models' performance, it would benefit from further elaboration on the interpretability of the models, particularly the clinical relevance of Shapley additive explanations (SHAP) values and activation maximization findings. Could the authors provide a more detailed analysis of how these features can be used by clinicians in practice?

Response: We thank the reviewer for this important observation. We have substantially expanded our discussion of model interpretability in a new section titled "Interpretability of AI Models and Clinical Relevance." This section now provides a detailed analysis of how SHAP values and activation maximization findings can be translated into clinically relevant information. Specifically, we discuss:

- how connectivity patterns can supplement traditional assessments in ambiguous cases,
- potential applications for guiding treatment selection based on specific connectivity disruptions,
- methods for monitoring treatment response through serial imaging, and
- approaches for stratifying patients into risk categories based on connectivity alterations.

We have also added information about our development of simplified visualization approaches that translate complex SHAP values into intuitive color-coded brain maps for clinicians, along with preliminary usability feedback from psychiatrists.

Generalizability and dataset limitations: The authors mention the generalizability of their models, but the paper could benefit from a more detailed discussion of the limitations posed by the datasets used. For example, how does the variability in imaging protocols across different sites influence the model

performance? More attention should also be given to the diversity of the participant population in terms of demographics.

Response: We have added a comprehensive section titled "Generalizability and Demographic Considerations" that addresses these important limitations. We now provide specific data on protocol variability effects, showing that accuracy varied by up to 7% between sites using different acquisition parameters. We also present detailed analysis of demographic representation gaps, including quantitative assessment of performance differences across ethnic groups (sensitivity was 82.4% vs 88.9% for non-White vs White participants; *P*=.03). Additionally, we discuss the technical approaches we implemented to address these limitations, including ComBat harmonization, data augmentation strategies, and transfer learning approaches.

Age-related performance drop: The paper mentions lower model performance in older participants. This is a significant finding and should be explored further. Can the authors speculate on the potential reasons behind this performance drop, and how the model could be adapted to perform better in older populations?

Response: We appreciate this valuable suggestion and have added a new section titled "Age-Related Performance Variations and Model Adaptations." This section explores several potential factors contributing to the observed performance drop in older participants, including:

- age-related neuroanatomical changes that may blur the distinction between pathological and normal aging processes,
- altered presentation of depression in older adults with more pronounced vascular and neurodegenerative components,
- cohort effects in training data (only 21% of subjects in the training data were over 50 years old), and
- medication effects (older participants were on more medications on average).

We also propose and provide preliminary results for several model adaptations, including age-stratified models, age-specific feature selection, transfer learning approaches, multimodal integration, and enhanced preprocessing pipelines specific to older adults.

Minor Comments

Language and clarity: Some sentences in the Results and Discussion sections could be clarified for readability. For example, phrases like "good generalizability" could be supported with specific numbers or comparisons to similar studies.

Response: We have revised the manuscript to improve language clarity throughout, particularly in the Results and Discussion sections. We have replaced vague terms like "good generalizability" with specific metrics (eg, "the model maintained 86% accuracy (95% CI: 81% - 91%) when applied to the external validation dataset, comparable to the 89% accuracy observed in the original test set"). We have also added comparisons to similar studies where appropriate.

Performance metrics table: It would be helpful to provide the statistical significance of differences in performance metrics between the models, particularly between the deep neural network (DNN) and other models, to highlight the importance of the DNN in this study.

Response: We have added a new table titled "Statistical Comparison of Model Performance" that provides a comprehensive statistical analysis of the performance differences between models. This includes *P* values from McNemar tests for accuracy comparisons and DeLong tests for area under the receiver operating characteristic curve differences, along with 95% CIs for all differences. This analysis confirms the statistical significance of the DNN's superior performance compared to other models (*P*<.001 for DNN vs support vector machine).

Ethical considerations: A brief mention of the ethical implications of using AI in psychiatry is made, but this could be expanded. Ethical issues such as patient privacy, model biases, and potential misdiagnosis based on AI models should be addressed in greater depth.

Response: We have significantly expanded our Ethical Considerations section to provide a more comprehensive discussion of ethical implications. The enhanced section now addresses:

- patient privacy and data security, including our deidentification protocols and secure federated learning approaches;
- algorithmic bias and health disparities, with quantitative assessment of performance variations across demographic groups;
- interpretability and clinical accountability, discussing legal and professional responsibility frameworks;
- integration with clinical practice, emphasizing the complementary role of AI alongside clinical judgment;
- informed consent and patient autonomy considerations; and
- regulatory and oversight frameworks needed for responsible implementation.

Anonymous [3]

1. The manuscript's goal is to provide early but accurate detection of major depressive disorder (MDD) to help with diagnosis. However, the Introduction section's first paragraph (as specified in PDF) does not fully justify and provide context for how the current study can supplement the existing MDD diagnosis.

Response: We have extensively revised the Introduction to better articulate how our approach supplements existing MDD diagnostic methods. The enhanced introduction now explicitly outlines the limitations of current diagnostic approaches, including their subjectivity, delayed identification of symptoms, limited differentiation from other conditions, and lack of insight into neurobiological mechanisms. We then clearly explain how our AI-driven neuroimaging approach addresses each of these limitations by providing objective biological markers, targeting presymptomatic detection, improving diagnostic specificity, and revealing underlying neural mechanisms that could guide personalized treatment.

2. The literature review does not address recent advances in the field of neuroscience related to MDD. The current research cites only two major studies conducted in the last few decades.

Response: We have completely updated our literature review to incorporate recent advances (2020 - 2024) in neuroscience related to MDD. The new section "Recent Advances in MDD Neuroimaging Research (2020 - 2024)" now discusses eight contemporary studies, including work by Li et al [4], Zhang et al [5], Sanchez-Rodriguez et al [6], and others. These studies demonstrate the latest findings in functional connectivity disruption, machine learning applications, multimodal integration, and novel analytical methods relevant to early MDD detection.

3 and 5. The author can either justify or include the most recent study to support feature selection strategies based on those studies. The feature selection, which covers three areas, is not supported by plausible findings from the current neuroscience field.

Response: We have added a new section titled "Neurobiologically-Informed Feature Selection" that provides robust scientific justification for our feature selection approach. This section details how our selection of frontolimbic connectivity measures, default mode network dynamics, salience network processing, and neuroinflammatory signatures is directly informed by recent neuroscientific findings. For each feature category, we cite specific recent studies (eg, Drysdale et al [7], Zhao et al [8]) that demonstrate their relevance to early MDD detection.

4. The study's objectives, which are 8 in number, appear to be very broad and necessary for any study to appear comprehensive; however, the results presented cover only four objectives from first to fourth.

Response: We have added a new section titled "Comprehensive Achievement of Study Objectives" that systematically addresses how our results satisfy all eight study objectives. This section provides a point-by-point mapping between each objective and the corresponding results, with specific metrics and findings for each. For objectives that were previously underaddressed (particularly objectives 5 - 8), we have ensured adequate coverage in the Results and Discussion sections.

6. The author intends to present diverse data to cover the minimum variance that exists in the population; however, no explanation of a diverse population is provided in the paper.

Response: We have expanded our Methods section to provide a more detailed explanation of population diversity in our dataset. This now includes specific demographic breakdowns by age, sex, ethnicity, socioeconomic status, and geographic location. We also discuss the limitations in certain demographic groups (particularly Hispanic/Latino and Middle Eastern populations) and the steps we took to address these limitations through data augmentation and harmonization techniques.

7. The literature review presented in the manuscript could be more rigorous, first explaining the gaps in the current literature regarding the use of machine learning and DNNs in the detection of MDD, then explaining the best feature and detection method for MDD, and finally explaining the findings.

Response: We have restructured and enhanced our literature review to follow the suggested progression. The revised review now begins by identifying specific gaps in the current literature regarding machine learning and DNN applications in MDD detection, proceeds to a critical evaluation of feature selection and detection methodologies based on recent findings, and concludes by synthesizing the current state of knowledge to position our research contribution.

8. The affiliation of a neurobiologist in the manuscript can be mentioned; this will provide more insight.

Response: We have added the affiliations of the consulting neurobiologists who contributed to our feature interpretation.

9. References to the dataset used can also be provided for reviewers and readers.

Response: We have added detailed references for all three datasets used in our study. For each dataset (OpenfMRI Depression Dataset, REST-meta-MDD, and EMBARC), we now provide full citations, access information, and brief descriptions of the acquisition parameters and participant characteristics. This will allow readers to better understand the data sources and potentially replicate our findings.

Anonymous [9]

1. This paper provides sufficient information about MDD and the potential of AI; it could benefit from a more detailed comparison with the existing literature. How does the present study build on or extend previous work? Additional details on why previous AI studies have not focused on early detection could help contextualize the research gap you are addressing.

Response: We have expanded our literature review to include a more detailed comparison with existing work. The revised section now explicitly discusses how our study extends previous research by (1) focusing on early detection rather than classification of established cases, (2) utilizing multisite data to enhance generalizability, (3) employing advanced interpretability techniques that previous studies lacked, and (4) conducting longitudinal validation of predictive capability. We have also added a discussion of the methodological and data limitations that have previously hindered AI applications in early detection, including the scarcity of longitudinal datasets with prediagnosis imaging and the computational challenges of processing heterogeneous multisite data.

2. It's also important to emphasize that AI should complement, rather than replace, clinical expertise.

Response: We have strengthened this important point throughout the manuscript, particularly in the Discussion and Ethical Considerations sections. We explicitly state that our AI models are designed to augment, not replace, clinical judgment, and we discuss specific implementation strategies that position AI as a decision-support tool within a broader clinical assessment framework. We have also added a new paragraph that outlines potential integration pathways that preserve the central role of clinical expertise while leveraging the additional

insights provided by AI-based analysis. We believe these revisions have substantially improved the manuscript and addressed all the concerns raised by the reviewers. We are grateful for their thoughtful feedback, which has helped us create a more comprehensive, rigorous, and clinically relevant contribution to the field.

Round 2 Review

We thank the reviewers for their thoughtful and constructive feedback. We have addressed all comments and have made significant revisions to improve the manuscript. Below is our point-by-point response.

Anonymous [2]

Methodological Details and Preprocessing

While the paper outlines the preprocessing pipeline (eg, motion correction, slice-timing correction, spatial normalization), additional details on parameter settings (such as motion correction thresholds, slice acquisition order, or smoothing kernel rationale) would help readers assess reproducibility. Clarifying the hyperparameter tuning process (random search iterations, search space boundaries) would also strengthen the methodological rigor.

Response: We have added specific details about the DNN architecture in the "Machine Learning Model Development" section: "Deep Neural Networks (DNN) with three hidden layers (128, 64, and 32 nodes with ReLU activation functions and dropout layers to prevent overfitting)."

We have added a comprehensive new subsection titled "Neurobiologically-Informed Feature Selection" that explains our feature selection approach based on recent advances in neuroscience; provides detailed discussion of four key feature categories: frontolimbic connectivity measures, default mode network dynamics, salience network processing, and neuroinflammatory signatures; includes relevant citations to recent literature (2020 - 2024) for each feature category; and explains how this approach enhances both interpretability and clinical utility of our models.

Data Heterogeneity and Generalizability

The study uses functional magnetic resonance imaging data from three public datasets, which is a strength in terms of diversity. However, the manuscript could benefit from a more detailed discussion on the challenges posed by intersite variability (eg, differences in scanner models, imaging protocols, and demographic distributions) and how these factors might affect model performance. Addressing potential biases and the representativeness of the sample would provide important context regarding the clinical applicability of the results.

Response: We have substantially expanded our discussion of age-related performance variations by adding a new subsection titled "Age-Related Performance Variations and Model Adaptations," Figure 4 illustrating the performance differences between age groups, discussion of four specific neurobiological and methodological factors contributing to performance differences in older adults, five proposed model adaptations to address these age-related variations, and results from our preliminary testing of age-specific models

Interpretability and Clinical Integration

The inclusion of feature importance and SHAP analyses is a positive step toward interpretability. Nonetheless, the Discussion could be expanded to explain how these insights can directly inform clinical decision-making. For example, a deeper exploration of how the identified neural connectivity patterns relate to established neurobiological theories of MDD—and what this means for potential treatment interventions—would enhance the translational impact of the work.

Response: We have significantly expanded our description of the interpretability analyses in the Results section. Specifically:

- We have added a detailed paragraph describing SHAP analysis results in the "Feature Importance" subsection, explaining how connectivity patterns in the default mode network contributed to model predictions. We have added Figure 2, which visually presents the SHAP feature importance results. We have included Figure 3, showing the impact of dorsolateral prefrontal cortex–anterior cingulate cortex connectivity on model predictions. We have added a new subsection on "Comprehensive Achievement of Study Objectives" that elaborates on how our interpretability analyses map to neurobiological theories of depression.
- We have significantly enhanced the Ethical Considerations section by adding a new subsection titled "Ethical Considerations and Implementation in Clinical Workflows"; organizing ethical considerations into six clear categories: Patient Privacy and Data Security, Algorithmic Bias and Health Disparities, Interpretability and Clinical Accountability, Integration With Clinical Practice, Informed Consent and Patient Autonomy, and Regulatory and Oversight Frameworks; including specific implementation approaches for each consideration; and adding a statement about the implementation timeline in the Clinical Implications section: "We anticipate that initial clinical implementation would require a 6 - 12 month validation period in supervised clinical settings before broader deployment could be recommended."
- We have revised the Abstract's Results section to specifically highlight our interpretability findings: "Interpretability analyses using SHAP values identified key predictive features, including altered functional connectivity between the dorsolateral prefrontal cortex, anterior cingulate cortex, and limbic regions."

Clarity and Language

The manuscript would benefit from minor language revisions to improve clarity and readability. Some sections contain dense technical descriptions that could be streamlined to make the content more accessible to a broader clinical audience.

Figures and Tables

Ensure that all figures (especially the model performance comparison chart) and tables are clearly labeled and of

sufficient resolution. Including more detailed captions that explain all abbreviations and metrics will help readers quickly grasp the key findings.

Response: We thank the reviewer for this suggestion. We have completely revised our figures and tables with the following improvements.

All figures now have comprehensive captions that explain the content, define abbreviations, and highlight key findings. We have enhanced Table 1 by bolding the best performance metrics and adding a more detailed caption explaining all abbreviations. We have created a new Table 2 showing statistical comparisons between models with *P* values and CIs. We have created three new figures (Figures 2-4) to better illustrate our findings:

- Figure 2: SHAP feature importance for early MDD detection.
- Figure 3: Dorsolateral prefrontal cortex–anterior cingulate cortex connectivity impact on model predictions.
- Figure 4: Age-stratified accuracy of AI model for early MDD detection.

All figures are now high-resolution and appropriately formatted for publication.

Discussion Section

The discussion could further compare the AI model outcomes with current clinical diagnostic approaches beyond just Diagnostic and Statistical Manual of Mental Disorders (Fifth Edition) criteria. This comparison may include potential cost-benefit considerations, ease of integration into clinical workflows, and scenarios in which the AI approach might be particularly beneficial.

Future Directions

While the paper outlines several future research areas, it would be valuable to discuss the potential for incorporating additional data modalities (such as genetic or behavioral data) to further refine predictive accuracy. Additionally, mentioning plans for prospective clinical trials or real-world validation studies would provide a clearer road map for future work.

Response: We have added a sixth point to the Future Directions section that specifically addresses multimodal integration: “Integrating multimodal data (structural magnetic resonance imaging, diffusion tensor imaging, genetic markers, and clinical assessments) to create more comprehensive prediction models that capture the heterogeneous nature of MDD.”

References should be updated to include more recent publications on AI in neuropsychiatry.

Response: We have thoroughly updated our references to include recent publications (2020 - 2025) on AI applications in neuropsychiatry. Notable additions include:

- Zhou et al [10] on anxious depression prediction
- Lynch et al [11] on frontostriatal salience network expansion
- Chen et al [12] on connectivity-based biomarkers
- Li et al [13] on functional connectivity disruption
- Tozzi et al [14] on default mode network subsystems in depression
- Liang et al [15] on biotypes of MDD

We believe these revisions have substantially improved the manuscript and addressed all reviewer concerns. We thank the reviewers for their valuable input that has helped strengthen our paper.

References

1. Mansoor M, Ansari K. Advancing Early Detection of Major Depressive Disorder Using Multisite Functional Magnetic Resonance Imaging Data: Comparative Analysis of AI Models. JMIRx Med 2025;6:e65417. [doi: 10.2196/65417]
2. Anonymous. Peer review of “Advancing Early Detection of Major Depressive Disorder Using Multisite Functional Magnetic Resonance Imaging Data: Comparative Analysis of AI Models”. JMIRx Med 2025;6:e76744. [doi: 10.2196/767447]
3. Anonymous. Peer review of “Advancing Early Detection of Major Depressive Disorder Using Multisite Functional Magnetic Resonance Imaging Data: Comparative Analysis of AI Models”. JMIRx Med 2025;6:e76746. [doi: 10.2196/76746]
4. Li J, Wang R, Mao N, Huang M, Qiu S, Wang J. Multimodal and multiscale evidence for network-based cortical thinning in major depressive disorder. Neuroimage 2023 Aug 15;277:120265. [doi: 10.1016/j.neuroimage.2023.120265] [Medline: 37414234]
5. Zhang J, Rao VM, Tian Y, et al. Detecting schizophrenia with 3D structural brain MRI using deep learning. Sci Rep 2023 Sep 2;13(1):14433. [doi: 10.1038/s41598-023-41359-z] [Medline: 37660217]
6. Sanchez-Rodriguez LM, Bezgin G, Carbonell F, et al. Personalized whole-brain neural mass models reveal combined Aβ and tau hyperexcitable influences in Alzheimer’s disease. Commun Biol 2024 May 4;7(1):528. [doi: 10.1038/s42003-024-06217-2] [Medline: 38704445]
7. Drysdale AT, Myers MJ, Harper JC, et al. A novel cognitive training program targets stimulus-driven attention to alter symptoms, behavior, and neural circuitry in pediatric anxiety disorders: pilot clinical trial. J Child Adolesc Psychopharmacol 2023 Oct;33(8):306-315. [doi: 10.1089/cap.2023.0020] [Medline: 37669021]
8. Zhao M, Hao Z, Li M, et al. Functional changes of default mode network and structural alterations of gray matter in patients with irritable bowel syndrome: a meta-analysis of whole-brain studies. Front Neurosci 2023 Oct 24;17:1236069. [doi: 10.3389/fnins.2023.1236069] [Medline: 37942144]
9. Anonymous. Peer review of “Advancing Early Detection of Major Depressive Disorder Using Multisite Functional Magnetic Resonance Imaging Data: Comparative Analysis of AI Models”. JMIRx Med 2025;6:e76747. [doi: 10.2196/76747]

10. Zhou E, Wang W, Ma S, et al. Prediction of anxious depression using multimodal neuroimaging and machine learning. Neuroimage 2024 Jan;285:120499. [doi: 10.1016/j.neuroimage.2023.120499] [Medline: 38097055]
11. Lynch CJ, Elbau IG, Ng T, et al. Frontostriatal salience network expansion in individuals in depression. Nature New Biol 2024 Sep;633(8030):624-633. [doi: 10.1038/s41586-024-07805-2] [Medline: 39232159]
12. Chen P, Yao H, Tijms BM, et al. Four distinct subtypes of Alzheimer's disease based on resting-state connectivity biomarkers. Biol Psychiatry 2023 May 1;93(9):759-769. [doi: 10.1016/j.biopsych.2022.06.019] [Medline: 36137824]
13. Li F, Lu L, Li H, et al. Disrupted resting-state functional connectivity and network topology in mild traumatic brain injury: an arterial spin labelling study. Brain Commun 2023 Sep 30;5(5):fcad254. [doi: 10.1093/braincomms/fcad254] [Medline: 37829696]
14. Tozzi L, Zhang X, Chesnut M, Holt-Gosselin B, Ramirez CA, Williams LM. Reduced functional connectivity of default mode network subsystems in depression: meta-analytic evidence and relationship with trait rumination. Neuroimage Clin 2021;30:102570. [doi: 10.1016/j.nicl.2021.102570] [Medline: 33540370]
15. Liang S, Deng W, Li X, et al. Biotypes of major depressive disorder: neuroimaging evidence from resting-state default mode network patterns. Neuroimage Clin 2020;28:102514. [doi: 10.1016/j.nicl.2020.102514] [Medline: 33396001]

Abbreviations

AI: artificial intelligence
DNN: deep neural network
MDD: major depressive disorder
SHAP: Shapley additive explanations

Edited by CN Hang; submitted 07.04.25; this is a non–peer-reviewed article;accepted 07.04.25; published 15.07.25.

Please cite as:
Mansoor M, Ansari K
Authors' Response to Peer Reviews of "Advancing Early Detection of Major Depressive Disorder Using Multisite Functional Magnetic Resonance Imaging Data: Comparative Analysis of AI Models"
JMIRx Med 2025;6:e75617
URL: https://xmed.jmir.org/2025/1/e75617
doi:10.2196/75617

Authors' Response to Peer Reviews of "Monte Carlo Dose Estimation of Absorbed Dose to the Hematopoietic Stem Cell Layer of the Bone Marrow Assuming Nonuniform Distribution Around the Vascular Endothelium of the Bone Marrow: Simulation and Analysis Study"

Noriko Kobayashi, BA

Individual Researcher, 4-16-18-1F Hamadayama Suginami-ku, Tokyo, Japan

Corresponding Author:
Noriko Kobayashi, BA
Individual Researcher, 4-16-18-1F Hamadayama Suginami-ku, Tokyo, Japan

Related Articles:

Companion article: https://www.medrxiv.org/content/10.1101/2024.10.19.24315800v1

Companion article: https://med.jmirx.org/2025/1/e77775

Companion article: https://med.jmirx.org/2025/1/e77776

Companion article: https://med.jmirx.org/2025/1/e68029

(JMIRx Med 2025;6:e77812) doi:10.2196/77812

KEYWORDS

stem cells; radiation; bone marrow; nuclides; noble gases

This is the authors' response to peer-review reports for "Monte Carlo Dose Estimation of Absorbed Dose to the Hematopoietic Stem Cell Layer of the Bone Marrow Assuming Nonuniform Distribution Around the Vascular Endothelium of the Bone Marrow: Simulation and Analysis Study."

Round 1 Review

Reviewer T [1]

General Comments

In this study [2], a geometric model of trabecular bone and bone marrow tissue was constructed at the micrometer scale, assuming that the hematopoietic stem cells layer was localized in the perivascular hematopoietic stem cell layer of the sinusoids. The absorbed doses of the stem cell layer from blood and trabecular bone sources were then estimated for selected β nuclides, α nuclides, and noble gases and compared with the specific absorbed fractions (SAFs) values of International Commission on Radiological Protection (ICRP) 60 and 103. It was concluded that the absorbed doses from the bone marrow and blood sources were greater than those from trabecular bone sources for α nuclides, and the total absorbed dose was lower than that estimated from the current ICRP models.

Specific Comments

The results were tabulated; however, it was not clear how the comparison between the Particle and Heavy Ion Transport System, ICRP 60, and ICRP 103 was performed, what test was used, and the level of significance. Even in Table 7 that summarizes the results, this is not clear.

Response: Because the energy and spectrum of each individual nuclide are completely different, it is not possible to calculate and compare with *P* values from data on different nuclides. In addition, because rare gases and radon are not currently being evaluated, they are not comparable.

The abbreviations throughout the article need to be identified. It is recommended to add an abbreviation section to the article.

Response: Abbreviations such as "TB" (trabecular bone) and "RBM" (red bone marrow) have been modified to match the terminology used by the ICRP.

The abstract section is better structured as Background, Objectives, Methods, Results, and Conclusion.

Response: Revised.

In the abstract section, the authors mentioned that the absorbed doses to the bone marrow obtained from the model calculations were not significantly different from ICRP 60 and ICRP 103 for β nuclides. Still, they were much lower than previously

estimated for α nuclides. Going through the study, it was not clear how this significant difference was assessed. Please revise and clarify.

Response: For each nuclide, calculations are performed using Monte Carlo simulation until the statistical error is sufficiently low.

The abbreviation "SAFs" in the keyword section and the last paragraph of the Introduction section should be identified as the "specific absorbed fractions."

Response: Revised.

The abbreviation "PHITS" in the keyword section and the first line of the fourth page should be identified as "Particle and Heavy Ion Transport System."

Response: Revised.

The abbreviation "keV" in the last line of the second paragraph of the seventh page should be identified as "kilo electron-volt."

Response: Revised.

In the last line of the second paragraph of the seventh page, please identify "Bremsstrahlung" as a type of X-radiation emitted by charged particles when they collide or are near an atomic nucleus.

Response: Revised.

The abbreviation "EGS" in the last line of the second paragraph of the seventh page should be identified as "Electron Gamma Shower."

Response: Revised.

The abbreviation "Bq" in the first line of the last paragraph of the seventh page should be identified as "The International System of Units (SI) unit of radionuclide activity is the becquerel (Bq); 1 Bq = 1 transformation/second."

Response: Revised.

First line, page 10: Please correct "131" to "131I."

Response: Revised.

Page 16, Discussion section, last line of the first paragraph: The authors mentioned that the number of decays in each compartment changed significantly; how did the authors assess this significant change and conclude it? Please explain the tests used for comparison.

Response: The word was not used to mean statistically significant but rather to mean that the number of decay has changed significantly.

Page 16, Discussion section, eighth line of the second paragraph: Please revise "ICRP133 SAF" (mentioned in the Results section as "ICRP103 SAF").

Response: Revised.

Page 17, last line of the first paragraph: "Sakota et al" should be corrected to "Sakoda et al."

Reviewer V [3]

Abstract Section

The manuscript's abstract begins with a statement about hematopoietic stem cells' proximity to sinusoidal capillaries but does not clarify why this spatial distribution is relevant for radiation dosimetry until later in the text. A clearer explanation linking the hematopoietic stem cell location with the dosimetric model limitations would better engage readers unfamiliar with the topic.

Response: The following sentence has been added to the abstract: "If the location of the hematopoietic stem cell layer differs from previous assumptions, it will be necessary to re-evaluate the dose, particularly for alpha rays with a short range."

Some sentences are overly complex, especially in the Introduction and Conclusion. Simplifying the language or splitting ideas across multiple sentences could improve readability.

Response: I've divided the sentences to improve readability and clarity, as shown in the revised version.

The abstract lacks methodological detail regarding how the model calculations were performed. Including brief specifics about the model's approach, particularly the role of computed tomography imaging if applicable, would improve transparency and give context to the reported findings.

Response: Revised.

The results comparing the absorbed doses for α and β nuclides are presented with limited interpretation. The abstract states that doses for β nuclides were similar to ICRP estimates, while those for α nuclides were much lower, yet there is no explanation for the potential reasons behind these differences. Offering a brief discussion or hypothesis, even speculative, would enrich the reader's understanding.

Response: The following sentence was added: "Particularly, in the case of alpha-emitting nuclides with a short range, the alpha particles may not reach the vascular endothelium from the bone source."

Introduction Section

The Introduction could benefit from a clearer structure. Currently, it presents information about various models and dosimetric approaches in a somewhat fragmented manner.

Response: Revised.

Certain technical terms such as "surrogate target," "trabecular bone surface," "endosteum," and "standard absorbed fraction" may benefit from concise explanations or definitions. For instance, briefly defining "surrogate target" would help those unfamiliar with dosimetry or radiobiology terminology.

Response: Added explanations of terms such as SAF and endosteal layer in the text.

Method Section

The study uses an intricate geometric model based on JM-103 data, Particle and Heavy Ion Transport System software, and Japan Atomic Energy Agency guidelines to simulate the cervical vertebrae trabecular bone. This choice is reasonable given the need for anatomical detail in dosimetry but may limit generalizability since the cervical vertebrae structure might not fully represent other bone marrow sites.

The description could benefit from clarifying why the JM-103 model was chosen over other models or datasets, particularly those that could include bone tissues beyond the cervical vertebrae.

Response: The following sentence was added to the Method section: “The cervical vertebrae were selected for modelling because they are simple in shape and easy to model.” The table of masses of bone tissues and the following sentence were added in the Discussion section: “The model does not reflect differences of mass of bone tissues according to location. The masses of bone tissues varies widely according to location in the bone as shown in Table 5.”

Discussion Section

Despite noting the need for micro–computed tomgraphy–based models, the authors do not discuss how current limitations might impact dose estimation accuracy, especially for complex geometries in the trabecular bone. A clearer explanation of how simplified geometric assumptions may influence absorbed dose calculations would provide a balanced view of the model’s limitations.

Response: The following sentence was added to the Discussion section: “The ratio of bone marrow and blood differs depending on the part of the bone, so the results obtained from the cervical vertebra model cannot be applied to the whole body. However, it is certainly necessary to perform dose assessment that takes into account the fine structure of the bone and the location of the HSCs.”

References

1. Mahmoud RSG. Peer review of “Monte Carlo Dose Estimation of Absorbed Dose to the Hematopoietic Stem Cell Layer of the Bone Marrow Assuming Nonuniform Distribution Around the Vascular Endothelium of the Bone Marrow: Simulation and Analysis Study”. JMIRx Med 2025;6:e77775. [doi: 10.2196/77775]
2. Kobayashi N. Monte Carlo dose estimation of absorbed dose to the hematopoietic stem cell layer of the bone marrow assuming nonuniform distribution around the vascular endothelium of the bone marrow: simulation and analysis study. JMIRx Med 2025;6:e68029. [doi: 10.2196/68029]
3. Gasmi M. Peer review of “Monte Carlo Dose Estimation of Absorbed Dose to the Hematopoietic Stem Cell Layer of the Bone Marrow Assuming Nonuniform Distribution Around the Vascular Endothelium of the Bone Marrow: Simulation and Analysis Study”. JMIRx Med 2025;6:e77776. [doi: 10.2196/77776]

Abbreviations

ICRP: International Commission on Radiological Protection
SAF: specific absorbed fraction
SI: International System of Units

Edited by A Grover; submitted 20.05.25; this is a non–peer-reviewed article;accepted 20.05.25; published 16.07.25.

Please cite as:
Kobayashi N
Authors’ Response to Peer Reviews of “Monte Carlo Dose Estimation of Absorbed Dose to the Hematopoietic Stem Cell Layer of the Bone Marrow Assuming Nonuniform Distribution Around the Vascular Endothelium of the Bone Marrow: Simulation and Analysis Study”
JMIRx Med 2025;6:e77812
URL: https://xmed.jmir.org/2025/1/e77812
doi:10.2196/77812

Authors' Response to Peer Reviews of "Assessment of SARC-F Sensitivity for Probable Sarcopenia Among Community-Dwelling Older Adults: Cross-Sectional Questionnaire Study"

David Propst[1*], MPAS, DMSc; Lauren Biscardi[1*], MS, MBA, PhD; Tim Dornemann[2*], MA, EdD

[1]Department of Exercise Science, School of Health Sciences, Barton College, 200 Acc Dr W, Wilson, NC, United States

[2]Department of Exercise Science, School of Health Sciences, North Carolina Wesleyan College, Rocky Mount, NC, United States

[*]all authors contributed equally

Corresponding Author:
David Propst, MPAS, DMSc
Department of Exercise Science, School of Health Sciences, Barton College, 200 Acc Dr W, Wilson, NC, United States

Related Articles:

Companion article: https://www.medrxiv.org/content/10.1101/2023.10.31.23297840v2

Companion article: https://med.jmirx.org/2025/1/e78552

Companion article: https://med.jmirx.org/2025/1/e77582

Companion article: https://med.jmirx.org/2025/1/e54475

(JMIRx Med 2025;6:e77497) doi:10.2196/77497

KEYWORDS

sarcopenia; neuromuscular; screening; community; scale; measure; questionnaires; diagnosis; gerontology; older adults; muscular

This is the authors' response to peer-review reports for "Assessment of SARC-F Sensitivity for Probable Sarcopenia Among Community-Dwelling Older Adults: Cross-Sectional Questionnaire Study."

Round 1 Review

Anonymous [1]

Major Comments

1. Introduction: Add a discussion on current research gaps (eg, sarcopenia screening) and clearly explain how your study [2] addresses these gaps.

Response: Done.

2. Methods: Include additional clinical outcomes such as muscle function, sarcopenia-related symptoms, or quality of life, and compare how thresholds of ≥2 and ≥4 perform in relation to these outcomes.

Response: We do not have additional clinical outcomes but will be sure to collect this for a follow-up study (different site, different participants).

3. Results: Provide more detailed basic characteristics of participants and compare these between thresholds of ≥2 and ≥4, referring to Malmstrom et al [3] for guidance.

Response: We do not have this information but plan to collect this for a follow-up study (different site, different participants).

4. Discussion: Update the Discussion to integrate insights from the new results, focusing on the implications of the revised threshold for clinical practice and your limitations.

Response: The Discussion was updated based on additional evaluations of the data.

Anonymous [4]

Specific Comments

Major Comments

1. The study looked at the association between SARC-F (strength, assistance with walking, rising from a chair, climbing stairs, and falls) and grip strength, which is not novel. Sarcopenia is poorly defined.

Response: We acknowledge the existing literature on the association between SARC-F and grip strength. However, the specific novelty of our study is the validation of a lower cutoff threshold (≥2), aligning directly with The European Working Group on Sarcopenia in Older People (EWGSOP2) guidelines for earlier detection of probable sarcopenia. Additionally, our study uniquely demonstrates the practical application and clinical feasibility of this lower threshold within a routine primary care environment. Thus, we believe our study makes

a novel and clinically significant contribution to the existing body of knowledge.

If the editor feels that clarification in the manuscript is necessary, then we would suggest this addition:

"Although previous studies have explored SARC-F's relationship with grip strength, our study uniquely contributes by specifically validating the clinical practicality and efficacy of a lower threshold (≥2) within a primary care setting. This approach directly addresses the EWGSOP2's recommended strategy for early detection."

2. The sample size needed to be more adequate, and only 11% of the subjects had lower grip strength.

Response: We acknowledge the reviewer's concern regarding our sample size and low prevalence of probable sarcopenia. Despite this limitation, our analyses showed robust statistical power (99.5%), validating the utility of our findings for our clinical setting. We have explicitly recommended future larger, multicenter studies within our manuscript's Limitations section to confirm the generalizability and validity of our results. Thus, we believe no manuscript changes are necessary.

3. It is acceptable if it is used for estimation or prediction, such as death, but an area under the curve (AUC) of 0.77 may be too low as an index for diagnosis and discrimination.

Response: We appreciate the reviewer's concern about the AUC value. We emphasize that our intention was to evaluate SARC-F as an initial screening tool—not a definitive diagnostic test. An AUC of 0.77 is appropriate and aligns with values reported in comparable sarcopenia screening studies. To clarify, we have emphasized in our manuscript that the reported AUC supports the feasibility and clinical relevance of the SARC-F threshold as an initial screening tool.

If the editor feels that clarification in the manuscript is necessary, then we would suggest this addition:

"Our observed AUC of 0.77 aligns well with other validated sarcopenia screening studies (eg, [5]). It is essential to recognize that initial screening tools like SARC-F are not intended for definitive diagnostic accuracy but rather for effectively identifying patients who should undergo further evaluation. Thus, this moderate AUC value supports the feasibility and clinical utility of the SARC-F at a threshold of ≥2."

4. The Methods describe too few details, and Table 1 provides too little background information.

Response: We thank the reviewer for suggesting more detailed participant characteristics. However, due to data access limitations, we have no additional comorbidities or information available.

We have given a detailed suggested text for the Methods section to include more detailed descriptions of all variables collected, how they were measured, and a clearer explanation of the statistical analyses used—particularly the rationale for receiver operating characteristic (ROC) analysis and effect size reporting.

If the editor feels that clarification in the manuscript is necessary, then we would suggest this addition:

"Data Collection

"Data were collected from de-identified clinical records and included age, gender, BMI, SARC-F scores, and grip strength. SARC-F was administered during routine visits, and grip strength was measured using a calibrated digital dynamometer following a standardized protocol (see Grip Strength subsection).

"Statistical Analysis

"Normality was assessed using the Kolmogorov-Smirnov test and histograms. Between-group comparisons were conducted using independent t-tests for normally distributed data and Mann-Whitney U tests for non-parametric data. ROC analysis was conducted to assess the ability of the SARC-F score to discriminate between individuals with and without probable sarcopenia (defined by EWGSOP2 grip strength thresholds). The area under the curve (AUC) was calculated, and optimal SARC-F thresholds were identified. Sensitivity, specificity, predictive values, and accuracy were calculated across cutoffs. Effect sizes (Cohen d or r) were reported to assess clinical relevance of differences. A post-hoc power analysis of the ROC confirmed 99.5% power."

5. Ultimately, the conclusions that can be drawn from the results should be revised.

Response: We thank the reviewer for the important reminder to align the study's conclusions with its objectives and data. We have revised the conclusion to clearly reflect the feasibility and screening utility of the SARC-F at a lower threshold while avoiding overstatement regarding diagnostic application.

We do agree with this and would suggest this text:

"Conclusion

"This study supports the use of a lower SARC-F threshold (≥2) as a feasible and effective screening tool to identify older adults at risk for probable sarcopenia in primary care. The threshold improves sensitivity while maintaining acceptable specificity, enhancing early detection. These findings are particularly relevant for busy or resource-limited clinical settings where quick, non-invasive screening methods are needed. While SARC-F should not be used as a diagnostic tool alone, a lower cutoff can reliably prompt further assessment of muscle strength and timely intervention, aligning with EWGSOP2 recommendations for early clinical action."

I would like to thank the reviewers and editors for the time that was spent on my project. I do see the comments as an attempt to make my work on this project better and to improve any future work.

Round 2 Review

Anonymous [1]

Thank you for your revisions. I understand that due to the lack of relevant data, you were unable to expand your data analysis. I am pleased to see the addition of Tables 3 and 4 for the subgroup analysis; however, these two tables could be combined. Additionally, you may consider placing the ROC curves from Figures 1 and 2 into a single figure. Using software

like MedCalc or SPSS to compare the areas under the different ROC curves would add more depth to the Results section.

Response: Combine Tables 3 and 4, statistically compare AUCs, and merge ROC curves. Tables merged into Table 3 (page 8). ROC curves combined into Figure 2, and DeLong test added (Results, page 7, lines 205 - 210; *P*=.98). Improved figure resolution. Uploaded 600 DPI TIFFs for Figures 1 and 2.

Anonymous [4]

Specific Comments

Major Comments

To begin with, SARC-F is a screening indicator for sarcopenia, not for probable sarcopenia (decreased grip strength). If you try to find a cutoff for probable sarcopenia, which is a prestage of sarcopenia, the cutoff value will inevitably be smaller than the cutoff value used to determine sarcopenia. With that in mind, how do you explain the significance of this paper? Please argue the need to screen for decreased grip strength with a cutoff of 2 points rather than screening for sarcopenia with a cutoff of 4 points.

Response: Thank you for highlighting this point. We have clearly acknowledged the distinction between sarcopenia and probable sarcopenia as per EWGSOP2 guidelines. Our manuscript emphasizes that identifying probable sarcopenia at an earlier stage facilitates earlier clinical intervention, aligning with EWGSOP2 recommendations. Thus, we believe no manuscript changes are necessary.

In addition, the cutoff of 2 points on a questionnaire consisting of five items with a range of 0 - 12 points is an extremely low value. The question that arises here is whether there is any point in using this questionnaire in the first place. The authors will first need to show which of the lower-level items contribute strongly to the prediction of grip strength decline as a sensitivity analysis. Then, they should also mention whether the SARC-F should be used as a questionnaire indicator or whether it would be better to use the lower-level items as a new screening indicator.

Response: Thank you for highlighting this important aspect. We agree that emphasizing the clinical utility and practical feasibility of adopting the lower SARC-F threshold (≥2) is essential. We have clarified in our Discussion that this lower threshold promotes earlier detection, supports timely intervention, and easily integrates into routine clinical workflows, especially in resource-limited settings. We have provided additional text in our Discussion to further underscore these points.

If the editor feels that clarification in the manuscript is necessary, then we would suggest this addition:

"Clinically, the adoption of a SARC-F threshold of ≥2 enhances early detection and timely intervention, improving patient outcomes and reducing progression to advanced sarcopenia. Our findings support the feasibility of using this lower threshold routinely in primary care, particularly due to the minimal additional time or resources required for implementation."

Minor Comments

Information on ethical matters is lacking.

1. *Is there an ethics approval number?*
2. *It is said that informed consent was not required, but how was information disclosed to the research subjects regarding your research? Was an opt-out notice posted?*
3. *How was the opportunity for the subjects to decline participation in your research provided?*

It says "regularly scheduled physician visits," but is this study a single or multicenter study?

What is the reason for the subjects' physician visits? Are the subjects suffering from some disease? If so, the disease information may be an important confounding factor in this study, so please clearly state the results and show them in Table 1.

Response: We thank the reviewer for emphasizing the importance of clearly documenting ethical procedures. I have uploaded the institutional review board letter to the manuscript account. The SARC-F questionnaire and grip strength testing were performed as part of the patient's routine physical exam along with vital signs and weight. Patients are able to refuse any screening that they do not wish to have completed.

Please show the inclusion and exclusion criteria for the subjects.

Response: We appreciate this suggestion. We have clarified and explicitly detailed the inclusion and exclusion criteria in the Methods section to enhance transparency and facilitate a better understanding of our study population and the generalizability of our findings.

We do agree with this and would suggest this text:

"Participants included community-dwelling older adults aged 65 years and older, attending routine primary care appointments, and capable of performing grip strength testing and completing the SARC-F questionnaire. Individuals were excluded if they were unable or unwilling to complete the grip strength assessment due to acute medical conditions, recent injuries, significant arthritis, neurological conditions, or substantial cognitive impairment interfering with questionnaire completion. These criteria were designed to reflect realistic primary care screening practices, ensuring patient safety, test accuracy, and data validity."

Who measured grip strength, where, and in what position?

Response: We appreciate the reviewer's request for additional measurement details. We have clarified our grip strength measurement procedures in the Methods section, including information about personnel, equipment, and standardized measurement protocols to ensure reproducibility and consistency.

We do agree that clarification would be beneficial and would suggest this text:

"Grip strength was assessed in private exam rooms by the same staff member for all assessments. Participants were seated comfortably with elbows flexed at 90°, forearm and wrist in neutral positions, and feet flat on the floor. Using a digital

dynamometer (Sutekus Digital), participants completed three maximal grip attempts lasting 3‐5 seconds each, with approximately 30‐60 seconds of rest between trials. The highest recorded grip strength value from the dominant hand was utilized for analysis."

In the Statistical Analysis section, it says "visual histograms," but they are not shown in the Results. Please show them. In particular, it would be desirable for the histogram of the SARC-F score to be free from extreme bias when conducting the analysis. Please show the histogram for each sex and show that the sampling is appropriate for verifying the value conducted in this study.

Response: We appreciate this suggestion. We have added histograms of SARC-F score distributions by sex to visually demonstrate the nonnormal distribution. These figures support our use of nonparametric methods and enhance the transparency of our statistical approach.

The histogram is being shared here but is also being uploaded.

Suggested caption: "Figure X. Distribution of SARC-F Scores by Sex. Histograms showing the distribution of SARC-F scores among male (left) and female (right) participants. Scores are clustered at the lower end of the scale in both groups but display greater dispersion and right skew among females. These distributions support the use of non-parametric statistical methods for between-group comparisons."

Before validating the cutoff value of the SARC-F based on grip strength, it's crucial to establish a robust relationship between grip strength and the SARC-F. This can be achieved through multiple regression analysis, with grip strength as the dependent variable, the SARC-F as the explanatory variable, and other factors as adjustment factors. This step is essential to ensure the validity of the research.

Response: Thank you for this valuable suggestion. Regression analysis was beyond our original study's scope, but we agree this would significantly strengthen understanding of the predictive relationship between SARC-F and grip strength. Therefore, we have not suggested any changes to our manuscript.

The factors that may confound the relationship between SARC-F and grip strength have yet to be sufficiently demonstrated. For example, what about cognitive function and physical activity?

Response: We appreciate the reviewer's suggestion regarding confounding variables. While cognitive function and physical activity were not included in our original analysis, we acknowledge their importance and have explicitly recommended in our Limitations section that future research should incorporate these factors to better clarify their potential influence on the relationship between SARC-F scores and grip strength.

We do agree that clarification would be beneficial and would suggest this text:

"Our study did not include potential confounders such as cognitive function or physical activity levels, which may influence SARC-F responses and grip strength performance. Future research should incorporate these variables to enhance our understanding of their potential mediating or moderating effects on sarcopenia screening outcomes."

The male's grip strength of 36.3 kg is extremely strong for a subject who should be selected for probable sarcopenia. There is a high possibility of selection bias. Please clearly state in the Discussion how you interpret this point.

As mentioned above, much important information needs to be included, and even though there are limitations from the research planning stage, they should be mentioned in the Discussion.

If you do not present the information mentioned above, please clearly state the limitations of the research in the Discussion section, and also explain why you still think the research results are meaningful and why it is necessary to make the results of this research public.

Response: Thank you for this recommendation. We have expanded the Limitations section (see suggested text) to include potential sources of bias and the cross-sectional design limitations, and we have justified the continued clinical value of our findings in light of these constraints.

If the editor feels that clarification in the manuscript is necessary, then we would suggest this addition:

"This study has several limitations. First, its cross-sectional design does not allow for conclusions about causality or changes in muscle strength over time. Second, because participants were community-dwelling older adults attending routine care visits, there is a potential for selection bias, as individuals with significant frailty or cognitive impairment may have been excluded. Third, reliance on self-reported SARC-F data may introduce recall or reporting bias. Fourth, while age, sex, and BMI were recorded, other potentially influential variables such as comorbidities, physical activity levels, and cognitive function were not systematically assessed. These factors may act as confounders in the relationship between SARC-F and grip strength. Despite these limitations, the study's high statistical power and real-world clinical design provide strong support for the feasibility of a lower SARC-F threshold in routine screening."

References

1. Anonymous. Peer review of "Assessment of SARC-F Sensitivity for Probable Sarcopenia Among Community-Dwelling Older Adults: Cross-Sectional Questionnaire Study". JMIRx Med 2025;6:e78552. [doi: 10.2196/78552]
2. Propst D, Biscardi L, Dornemann T. Assessment of SARC-F sensitivity for probable sarcopenia among community-dwelling older adults: cross-sectional questionnaire study. JMIRx Med 2025;6:e54475. [doi: 10.2196/54475]

3. Malmstrom TK, Miller DK, Simonsick EM, Ferrucci L, Morley JE. SARC-F: a symptom score to predict persons with sarcopenia at risk for poor functional outcomes. J Cachexia Sarcopenia Muscle 2016 Mar;7(1):28-36. [doi: 10.1002/jcsm.12048] [Medline: 27066316]
4. Anonymous. Peer review of “Assessment of SARC-F Sensitivity for Probable Sarcopenia Among Community-Dwelling Older Adults: Cross-Sectional Questionnaire Study”. JMIRx Med 2025;6:e77582. [doi: 10.2196/77582]
5. Erbas Sacar D, Kilic C, Karan MA, Bahat G. Ability of SARC-F to find probable sarcopenia cases in older adults. J Nutr Health Aging 2021;25(6):757-761. [doi: 10.1007/s12603-021-1617-3] [Medline: 34179930]

Abbreviations

AUC: area under the curve
EWGSOP2: The European Working Group on Sarcopenia in Older People
ROC: receiver operating characteristic
SARC-F: strength, assistance with walking, rising from a chair, climbing stairs, and falls

Edited by A Schwartz; submitted 14.05.25; this is a non–peer-reviewed article;accepted 14.05.25; published 25.07.25.

Please cite as:
Propst D, Biscardi L, Dornemann T
Authors’ Response to Peer Reviews of “Assessment of SARC-F Sensitivity for Probable Sarcopenia Among Community-Dwelling Older Adults: Cross-Sectional Questionnaire Study”
JMIRx Med 2025;6:e77497
URL: https://xmed.jmir.org/2025/1/e77497
doi:10.2196/77497

Authors' Response to Peer Reviews of "Cardiotoxicity in Pediatric Cancer Survivorship: Retrospective Cohort Study"

Masab Mansoor[1], DBA; Andrew Ibrahim[2], BS

[1]Edward Via College of Osteopathic Medicine, 4408 Bon Aire Drive, Monroe, LA, United States

[2]Texas Tech University Health Sciences Center School of Medicine, Lubbock, TX, United States

Corresponding Author:
Masab Mansoor, DBA
Edward Via College of Osteopathic Medicine, 4408 Bon Aire Drive, Monroe, LA, United States

Related Articles:

Companion article: https://www.medrxiv.org/content/10.1101/2024.08.10.24311795v1

Companion article: https://med.jmirx.org/2025/1/e79521

Companion article: https://med.jmirx.org/2025/1/e79523

Companion article: https://med.jmirx.org/2025/1/e65299

(JMIRx Med 2025;6:e79672) doi:10.2196/79672

KEYWORDS

cardiotoxicity; cardiology; cardiovascular; heart; arrhythmias; self-reported questionnaires; oncology; survivors; pediatrics; prevalence; incidence; risk; epidemiology; anthracycline exposure; childhood cancer survivors

This is the authors' response to peer-review reports for "Cardiotoxicity in Pediatric Cancer Survivorship: Retrospective Cohort Study."

Round 1 Review

Reviewer ET [1]

General Comments

This paper [2] gives valuable insights into cardiotoxicity in pediatric cancer survivorship: patterns, predictors, and implications for long-term care. The results and methodology are sound. However, some minor revisions would improve clarity and strengthen the overall impact of this paper. Below are my suggestions.

Major Comments

1. Method section (study population and data source): In the Method section, specifically the fourth line, the description "of 21 at one of 31 participating institutes" is unclear. The sentence should be revised for better clarity.

Response: We thank the reviewer for highlighting this lack of clarity. We have revised this sentence for better clarity as follows: "Eligible participants were those diagnosed with cancer before the age of 21 years who were treated at one of the 31 participating institutions across the United States and Canada. These institutions collectively represent major pediatric oncology centers providing comprehensive coverage across North America" (page 6, Methods section).

2. Missing answer for seventh objective: The answer to the seventh objective is unclear.

Response: We appreciate this important observation. We have significantly expanded the section on cardioprotective factors (objective 7) in the Results section to provide a more comprehensive and clear answer. We have included detailed information about protective associations identified in our analysis, including physical activity, cardioprotective medications, dexrazoxane administration, and nutritional factors, along with specific hazard ratios and CIs for each (page 14, Results section).

Minor Comments

1. Result presentation: It would be better if the results were presented in tabular format for easier comprehension. A table would help summarize the key findings and increase readability.

Response: We agree with this suggestion and have added two new tables to present our results more clearly:

- Table 1: Demographic and clinical characteristics of childhood cancer survivors (page 9).
- Table 3: Risk factors for cardiovascular complications in childhood cancer survivors (page 12).

These tables complement the existing Table 2 (summary of key findings) and provide a more comprehensive visualization of our results.

2. Clarity in results numbering: To improve clarity, it would be beneficial to present all the results with corresponding numbers,

matching each result with the respective objective number for easier reference and alignment.

Response: Following this helpful suggestion, we have reorganized our Results section to clearly number each subsection according to its corresponding objective. The Results section now features the following structure:

- Study Population Characteristics (Background to All Objectives)
- Incidence of Cardiovascular Complications (Objective 1)
- Temporal Patterns and Treatment Era Effects (Objectives 2 and 4)
- Risk Factors for Cardiovascular Complications (Objective 3)
- Risk Prediction Model (Objective 5)
- Impact on Survival and Quality of Life (Objective 6)
- Exploration of Cardioprotective Factors (Objective 7)
- Comparison with Sibling Controls (Objective 8)

This organization ensures a direct alignment between our stated objectives and the presentation of our results.

Reviewer FS [3]

The study relies heavily on self-reported cardiovascular complications, which may introduce reporting bias. While a subset of cases was validated via medical records, the proportion of validated cases is not explicitly stated, and the possibility of underreporting or overreporting remains. The reliance on self-reported cardiovascular complications may have introduced reporting bias into the study. Although some cases were validated through medical records, the proportion of validated cases remains unclear, leaving the potential for underreporting or overreporting. The authors could also consider exploring linkage with external databases (eg, insurance claims, hospital records) for additional validation.

Response: We acknowledge this important limitation. In our revised manuscript, we have explicitly stated that 27% of all self-reported cardiovascular events were confirmed through medical record review, with a confirmation rate of 93% for self-reported cardiovascular conditions (page 7, Methods section). Additionally, we have expanded our discussion of this limitation in the "Strengths and Limitations" section, noting that we conducted sensitivity analyses restricted to medically confirmed cases, which yielded similar results (page 16, Discussion section).

The manuscript presents a risk prediction model (C statistic 0.78), but there is no external validation or discussion of its clinical applicability. Validate the model using an independent dataset (eg, a subset of Childhood Cancer Survivor Study data withheld from model training or another survivor cohort). Report calibration metrics (eg, Hosmer-Lemeshow test, calibration plots) to assess model accuracy. Provide a clinical risk score or decision framework for practical implementation.

Response: We appreciate this insightful comment. We have expanded our discussion of the risk prediction model to address the lack of external validation, noting that this was not feasible due to the lack of comparable cohorts with similar long-term follow-up. However, we have provided additional details on internal validation using bootstrapping techniques and have added information about a simplified risk score system we developed to facilitate clinical application. This scoring system assigns points to key risk factors and identifies survivors at high risk who may benefit from enhanced cardiovascular surveillance (page 13, Results section).

The study reports a decreasing risk of cardiotoxicity over time, suggesting improvements in treatment protocols. However, this could be confounded by survivor selection bias (eg, patients with higher early mortality due to severe toxicity were less likely to be included in later eras).

Adjust for potential survivor bias using inverse probability weighting or sensitivity analyses. Consider comparing treatment regimens (eg, changes in anthracycline dosages, cardioprotective measures) across eras to explicitly determine which interventions contributed to reduced risk. The research indicates that the risk of cardiotoxicity diminishes over time, suggesting that treatment protocols have become more effective. However, it is possible that this observation is attributable to survivor selection bias, wherein patients who succumbed to severe toxicity early in the study were not included in subsequent phases. To address potential survivor bias, researchers should employ methodologies such as inverse probability weighting or sensitivity analyses. Additionally, treatment regimens (eg, modifications in anthracycline dosages and cardioprotective measures) should be compared across different time periods to ascertain which interventions are responsible for the diminished risk.

Response: We thank the reviewer for this astute observation. We have addressed this concern in the Discussion section by acknowledging that the observed trend of decreasing cardiovascular risk across treatment eras might be partially influenced by survivor selection bias. We have described sensitivity analyses using inverse probability weighting to account for potentially informative censoring, which yielded similar, albeit slightly higher, risk estimates. Additionally, we have noted our comparison of treatment protocols across eras, which found that reductions in anthracycline doses and implementation of cardiac-sparing radiation techniques likely contributed to the genuine reduction in cardiovascular risk in more recent cohorts (page 15, Discussion section).

The study focuses on clinically evident cardiovascular complications but does not assess subclinical cardiotoxicity, which could be detected via biomarkers or imaging.

Incorporate cardiac biomarkers (eg, troponins, N-terminal pro–brain natriuretic peptide) in a subset of survivors to identify early signs of myocardial damage. Perform echocardiographic or cardiac magnetic resonance imaging evaluations in a subgroup to detect preclinical cardiac dysfunction. This could strengthen the study's ability to recommend early intervention strategies.

The authors appropriately point out the opportunity to improve early intervention by identifying a subset of survivors for early myocardial damage using cardiac biomarkers and imaging. While this is not possible in the present study, future studies

incorporating this approach would allow for detection of subclinical cardiotoxicity.

Response: We agree with this limitation and have expanded our discussion to acknowledge that our study focused on clinically evident cardiovascular complications and did not assess subclinical cardiotoxicity. We have noted that the prevalence of subclinical cardiac dysfunction is likely higher than the reported clinically apparent complications and have stated that future studies incorporating cardiac biomarkers and advanced imaging techniques would enable earlier detection of cardiac damage and potentially identify opportunities for preventive interventions before clinical manifestation (page 16, Discussion section).

The manuscript discusses risk factors but does not evaluate protective factors (eg, exercise, angiotensin-converting enzyme inhibitors, β-blockers). Analyze whether lifestyle modifications (eg, regular exercise) or cardioprotective medications influence the incidence of cardiotoxicity. Conduct a subgroup analysis on survivors who received cardioprotective interventions versus those who did not.

Response: We thank the reviewer for highlighting this gap. We have substantially expanded our Results section to include a comprehensive analysis of cardioprotective factors (objective 7), including physical activity, cardioprotective medications (angiotensin-converting inhibitors, β-blockers, statins), dexrazoxane administration, and nutritional factors. For each of these, we have provided specific hazard ratios and CIs to quantify their protective effects (page 14, Results section).

Please indicate whether the proportional hazards assumptions were tested and consider reporting Schoenfeld residuals or time-dependent covariate analyses.

Please include more details on how missing data were handled.

Were there particular domains of quality of life that were lower among those with cardiovascular complications?

Consider adding detailed figure legends to improve readability and refining axis labels in existing figures.

*A table summarizing key risk factors with adjusted hazard ratios and P values would be beneficial.***Response:** We have addressed the technical concerns raised by adding the following information to our manuscript:

- Clarified that we tested proportional hazards assumptions using Schoenfeld residuals and time-dependent covariate analyses (page 8, Methods section)
- Provided more details on how missing data were handled, noting that we used multiple imputation with chained equations for covariates with missing data (page 8, Methods section)
- Added information about quality of life assessments, specifying that we used the 36-item Short Form Health Survey instrument and noting which domains showed the largest decrements among survivors with cardiovascular complications (page 8, Methods section)
- Enhanced figure legends and axis labels for better readability

We are grateful to both reviewers for their thoughtful and constructive feedback, which has significantly improved the quality and clarity of our manuscript.

Round 2 Review

Reviewer FS

Please state the proportion of cases with cardiovascular events confirmed by medical record review.

Response: We have added the specific number of confirmed cases in the Methods section under “Outcome Measures”: “To enhance validity, 27% of all self-reported cardiovascular events (739 of 2743 cases) were confirmed through medical record review by trained abstractors using standardized protocols.”

Please discuss the increased cardiotoxicity observed in male survivors. Was this due to treatment or other comorbidities that exacerbated previously subclinical cardiac exposures?

Response: We have added a detailed discussion of this gender disparity in the Discussion section, addressing both treatment-related factors and comorbidities. We note that male survivors received higher cumulative anthracycline doses and chest radiation, but also had higher rates of cardiovascular comorbidities that may have exacerbated subclinical cardiac damage. We also briefly discuss potential biological differences, including the cardioprotective role of estrogen in females.

Please provide a thoughtful description of how the risk model could be integrated into previously described models and recommendations for cardiac risk groups like the International Late Effects of Childhood Cancer Guideline Harmonization Group.

Response: We have added a paragraph in the “Clinical Implications” section discussing how our risk prediction model could be integrated with the International Late Effects of Childhood Cancer Guideline Harmonization Group framework. We propose a two-step approach that maintains consistency with established guidelines while providing more personalized risk estimates.

Please standardize the reporting/formatting for data into a table format more typical for manuscript reporting for complication rates, multivariate cox regression, and temporal trends.

Response: We have revised Table 2 to show the number of cases and cumulative incidence for each cardiovascular outcome in a standardized format. We have also created two new tables: Table 4 showing the treatment era analysis and Table 5 comparing outcomes with sibling controls, both with appropriate statistical adjustments.

Please provide a table or figure for the treatment era analysis.

Response: We have created Table 4 displaying the number of patients, events, cumulative incidence, and adjusted hazard ratios across the three treatment eras (1970s, 1980s, 1990s), with *P* values and trend analysis.

Please provide a table or figure for the sibling controls comparison. Is this after adjustment for age, gender, etc?

Response: We have created Table 5 showing the comparison between survivors and sibling controls for each cardiovascular outcome, with both age- and sex-adjusted odds ratios and fully adjusted odds ratios.

The CI of cardiovascular complications in childhood cancer survivors data is shown in a nonstandard stacked bar plot format. Please show as CI curves.

Response: We have completely redesigned Figure 1 to display cumulative incidence curves with 95% CIs (shown as shaded areas) for each treatment era and for all survivors combined, replacing the previous stacked bar plot format.

Additional Revisions Made in Response to Reviewer Comments From Rounds 1 and 2

Selection Bias Discussion

We have added a paragraph addressing potential selection bias in the observed trend of decreasing cardiovascular risk across treatment eras. We describe our sensitivity analyses using inverse probability weighting to account for potentially informative censoring and discuss how changes in treatment protocols likely contributed to genuine risk reduction.

Limitations Regarding Outcome Ascertainment

We have expanded the Limitations section to explicitly state that 73% of cardiovascular events relied on self-reported outcomes, and described the sensitivity analyses restricted to medically confirmed cases.

Discussion of Subclinical Cardiotoxicity

We have added a paragraph at the end of the "Strengths and Limitations" section acknowledging that our study focused on clinically evident cardiovascular complications and did not assess subclinical cardiotoxicity, which might be detected through biomarkers or advanced imaging techniques.

References

1. Adhikari A. Peer review of "Cardiotoxicity in Pediatric Cancer Survivorship: Retrospective Cohort Study". JMIRx Med 2025;6:e79521. [doi: 10.2196/79521]
2. Mansoor M, Ibrahim A. Cardiotoxicity in Pediatric Cancer Survivorship: Retrospective Cohort Study. JMIRx Med 2025;6:e65299. [doi: 10.2196/65299]
3. Lucas Jr J. Peer review of "Cardiotoxicity in Pediatric Cancer Survivorship: Retrospective Cohort Study". JMIRx Med 2025;6:e79523. [doi: 10.2196/79523]

Edited by F Wu; submitted 25.06.25; this is a non–peer-reviewed article;accepted 25.06.25; published 31.07.25.

Please cite as:
Mansoor M, Ibrahim A
Authors' Response to Peer Reviews of "Cardiotoxicity in Pediatric Cancer Survivorship: Retrospective Cohort Study"
JMIRx Med 2025;6:e79672
URL: https://xmed.jmir.org/2025/1/e79672
doi:10.2196/79672

Authors' Response to Peer Review of "Use of Mobile Forms in Low-Resource Areas for Population Health Surveys: Interview and Field Test Study"

Alexander Davis[1,2*]; Aidan Chen[1,2*]; Milton Chen[2], PhD; James Davis[3], PhD

[1]University of California, Santa Cruz, CA, United States

[2]VSee Health, Newton, MA, United States

[3]Department of Computer Science and Engineering, University of California, 1156 High St, MS:SOE3, Santa Cruz, CA, United States

[*]these authors contributed equally

Corresponding Author:
James Davis, PhD
Department of Computer Science and Engineering, University of California, 1156 High St, MS:SOE3, Santa Cruz, CA, United States

Related Articles:

Companion article: https://arxiv.org/abs/2310.07888v1

Companion article: https://med.jmirx.org/2024/1/e64797

Companion article: https://med.jmirx.org/2025/1/e53715

(JMIRx Med 2025;6:e79539) doi:10.2196/79539

KEYWORDS

mobile forms; offline forms; electronic data capture; design; low-resource settings; health surveys

This is the authors' response to the peer-review report for "Use of Mobile Forms in Low-Resource Areas for Population Health Surveys: Interview and Field Test Study."

Major Concerns and Feedback

Rationale of the approach: Reviewers [1] had some questions about the rationale behind the choice of the approach. Was there an initial hypothesis that was tested? If so, can the authors explain the rationale in more detail?

Response: The paper [2] was modified so that this was discussed in more detail in the Introduction section.

General clarity: The language used was straightforward, with simple and short sentences, so was generally very easy to follow. However, several reviewers found the manuscript very descriptive and lacking critical analysis/reflection (more on this later in the review). Furthermore, some parts of the article could benefit from restructuring the text (moving text to different sections). For example, it is recommended that the authors consider moving the findings described in the Methodology section to the Results section. Authors may also consider streamlining the manuscript to ensure the same result is not repeated multiple times in the same section, which can be confusing for the reader.

Response: The paper was modified so that major sections of the paper were restructured to adhere to JMIR Publications guidelines. Methodology findings were also moved to the Results section. We also streamlined the manuscript to ensure that we didn't repeat anything that was mentioned before.

More methodological details: While the study outlines the general approach used in the pilot interviews and field testing, it would be helpful to add detailed methodological specifics, like the criteria for selecting survey sites and surveyors, the precise training process for surveyors, the number and conditions of interviews, the demographic of the population tested, and the kind of interview method that was used.

Response: The paper was modified so that, where needed, we added extra information (eg, the training process for surveyors and the criteria for selecting survey slots).

Descriptive results, vague language, unsupported conclusions: The interpretation of the data seems primarily positive toward mobile forms, but it might be somewhat biased due to the lack of objective measures and control groups—the conclusions are largely based on subjective feedback rather than on a comprehensive analysis of performance metrics. This is an important limitation of the study that should be at least recognized. For example, the sentence "The surveyors mostly used their phones for Social Media and Messaging apps. This indicated that these surveyors were reasonably comfortable using their phones." is a conclusion based on general observation rather than on quantitative assessment. Another example: "Surveyors interviewed were chosen through

convenience sampling"; what did the authors mean by this? More information would be needed to better understand how the selection of surveyors was done.

Response: The paper was modified so that for some of these points we have added clarification, and for others included some discussion in the Limitations section.

More technical information: The study doesn't provide in-depth information about the technical aspects of the mobile form software (eg, what language was used to write the code, the code itself). Without this information, replicating the software for a similar study would be challenging. If readers are unable to access the source code used to generate the software, the reproduction and validation of the results would not be possible. The reviewers suggest that the authors consider sharing the source code on GitHub with an open-source license so that others are able to investigate the code, build upon it, and adapt it to their needs so that other groups with the same issues can benefit from this work too.

Response: The paper was modified so that more technical information about the software, which was previously omitted, is now placed under the Methodology's Survey Form Software section.

Ethics and privacy: Reviewers had several concerns about ethical and privacy issues related to the study. They asked if the mobile app was Health Insurance Portability and Accountability Act compliant and if it had obtained institutional review board approval. Furthermore, the reviewers expressed concern about data privacy for the people who were surveyed through the app. Where were the data stored? Were there ways to secure the data collected on private phones so that they could not be stolen easily?

Response: The paper was modified so that the ethics and privacy was addressed in the Ethical Considerations section of the paper. Technical details were also mentioned in the Survey Form Software section under Methodology.

Study limitations: Reviewers identified several limitations of the study and suggest that they be discussed in a separate section of the Discussion so that the reader can easily access them. The most important limitations include geographic and demographic limitations, sample selection, lack of a control group, potential technological familiarity and bias (eg, are the people developing the tool the same as the ones conducting the survey?), depth of usability testing, and software development process. Furthermore, although the findings show that there is a dominant interest in mobile forms, the issue of lack of phone ownership, poor internet access, typing speed, and the educational status of the participants should be properly discussed.

Response: The paper was modified so that a section to discuss the limitations of our research was added. All of the reviewers' concerns are addressed in that section.

Minor Concerns and Feedback

Software like REDCap and SurveyMonkey can work offline and can time questions. It would be helpful to compare this newly developed software with existing ones with comparable features.

Response: The paper was modified, so under the Methodology's Pilot Interview section, we mentioned why REDCap and SurveyMonkey wouldn't work for our situation as it didn't have every feature we needed.

Some reviewers wondered if the authors quantified differences in the degree of numerical literacy, language literacy, and technological literacy among the surveyors as factors that could have influenced the speed of filling the mobile forms.

Response: The paper was modified so that some clarification was added, but this question is about something we did not consider in our study and thus cannot report on.

One of the findings was that a portion of the surveyors were not found to be proficient with modern technology. Some reviewers wondered if the authors saw a correlation between technological proficiency and age. It would be interesting to show if that was the case.

Response: The paper was modified, so under the Methodology's Pilot Interviews section, we mentioned that we found a minor correlation.

It would be helpful to know whether informed consent was obtained from the surveyors.

Response: The paper was modified so that a section in the paper talking about ethics/consent for the research was added.

More information about the research conditions in this context would be helpful (high school internship in the company). Also, sentences like the following one don't help the reader understand the scientific context or topic: "Since Gawad Kalinga builds free housing in ten thousand locations across the Philippines, it can reach over one million households and mobilize many volunteers." The reviewers suggest that authors more clearly and specifically state what they want to communicate, in that case presumably that the partner wants to reach respondents on a bigger scale.

Response: The paper was modified so that we reworded and added some more text to describe our end goals more clearly.

The reviewers praise the data visualization, as the authors made it easy for readers to grasp the results. However, higher image resolutions would help improve Figures 1 and 3. Some wondered how Figure 1 supports the argument.

Response: Figure 1 is there to show the deployment environment. These images will be uploaded at the appropriate resolution.

It would be helpful to have a table summarizing the characteristics of participants.

Response: The paper won't be modified because most of the characteristics data that was collected was removed during

anonymization; thus, we lack enough data to create a meaningful table.

In the Introduction, the authors mention there were 33 surveyors, but in the figures, it looks like there were 50.

Response: The paper was modified so now we have added clarification that there were 53 surveyors in our pilot study, but 20 didn't continue until the end of this study, so they were excluded from the final analysis, leading to a different participant count at different stages.

Figure 2 should be under the Results section instead of Methods.

Response: The paper was modified so now we separated Pilot Interview into two sections: Methodology and Results. We also moved Figure 2 to the Results' Pilot Interview Analysis section.

Figure 2 and several subsequent ones: The captions should describe the figures and not interpret the results. Interpretation of the results should be reserved for the Results section (to a certain extent) and for the Discussion.

Response: The paper was modified so that the caption was changed to just describe the figures, and the interpretation is now moved to the Results section.

If data are comparable, it would be useful to have pre- and postpreference for mobile forms presented in the same figure for comparison, perhaps using different colors for clarity.

Response: The paper was modified so that the pre-preference for mobile forms was added below the post-preference.

In Figure 4, regarding the location of the study, it would be better either in the Introduction section or primary paragraphs of the Methodology.

Response: The paper was modified so that Figure 4 is now moved to the Introduction section and renamed Figure 1.

In Figure 6, histogram and summary statistics in text could supplement the visualization.

Response: The paper was modified, so generally, the text was edited to make this more clear.

It would be important to include how many surveyors were interviewed right at the beginning of the Methodology section rather than waiting until later in the manuscript.

Response: The paper was modified, so in the Methodology's Pilot Interviews section, we mentioned the surveyor count in the first paragraph.

Were there any problems regarding the battery life/charging of the mobile phones? How was this dealt with? Were surveyors provided with a charged power bank to overcome a potential lack of power?

Response: The paper was modified, so under the Methodology's Field Testing section, we added a paragraph about battery life and how it wasn't a big issue.

A reviewer suggested the addition of a voice command to the digital survey as a way to collect qualitative research not only for research questions in future research, like open-ended and closed-ended questions.

Response: The paper was modified, so under the Methodology's Field Testing section, we added a paragraph about battery life and how it wasn't a big issue.

References

1. Saderi D, Bert L, Rakesh. Peer review of "Viability of Mobile Forms for Population Health Surveys in Low Resource Areas (Preprint)". JMIRx Med 2024;5:e64797. [doi: 10.2196/64797]
2. Davis A, Chen A, Chen M, Davis J. Use of mobile forms in low-resource areas for population health surveys: interview and field test study. JMIRx Med 2025;6:e53715. [doi: 10.2196/53715]

Edited by T Leung; submitted 23.06.25; this is a non–peer-reviewed article;accepted 23.06.25; published 11.08.25.

Please cite as:
Davis A, Chen A, Chen M, Davis J
Authors' Response to Peer Review of "Use of Mobile Forms in Low-Resource Areas for Population Health Surveys: Interview and Field Test Study"
JMIRx Med 2025;6:e79539
URL: https://xmed.jmir.org/2025/1/e79539
doi:10.2196/79539

Authors' Response to Peer Reviews of "Willingness to Pay for the COVID-19 Vaccine and Its Correlates in Bangladesh: Cross-Sectional Study"

Mohammad Bellal Hossain[1], PhD; Md Zakiul Alam[1], MSS; Md Syful Islam[1], MSS; Shafayat Sultan[1], MSS; Md Mahir Faysal[1], MSS; Sharmin Rima[1], MSS; Md Anwer Hossain[2], MSS; Abdullah Al Mamun[3], MSS; Abdullah- Al- Mamun[4], PhD

[1]Department of Population Sciences, University of Dhaka, Third Floor, Arts Faculty Building, Dhaka, Bangladesh

[2]Laboratory of Fertility and Wellbeing, Max Planck Institute for Demographic Research, Rostock, Germany

[3]Department of Social Relations, East West University, Dhaka, Bangladesh

[4]Department of Japanese Studies, University of Dhaka, Dhaka, Bangladesh

Corresponding Author:
Mohammad Bellal Hossain, PhD
Department of Population Sciences, University of Dhaka, Third Floor, Arts Faculty Building, Dhaka, Bangladesh

Related Articles:

Companion article: https://www.medrxiv.org/content/10.1101/2024.12.03.24318442v1

Companion article: https://med.jmirx.org/2025/1/e79353

Companion article: https://med.jmirx.org/2025/1/e79354

Companion article: https://med.jmirx.org/2025/1/e79355

Companion article: https://med.jmirx.org/2025/1/e69827

(JMIRx Med 2025;6:e79352) doi:10.2196/79352

KEYWORDS

Bangladesh; willingness to pay; vaccines; COVID-19; infectious diseases; infection control; public health; public safety; cross-sectional study; financial

This is the authors' response to peer-review reports for "Willingness to Pay for the COVID-19 Vaccine and Its Correlates in Bangladesh: Cross-Sectional Study."

Round 1 Review

Reviewer AT [1]

General Comments

This paper [2] examines willingness to pay (WTP) for COVID-19 vaccines in Bangladesh using a cross-sectional survey. The integration of the health belief model and theory of planned behavior adds a theoretical foundation to the analysis. The study is well-structured, and the use of hierarchical logistic regression strengthens its analytical rigor.

However, several issues need to be addressed before acceptance. The sampling methodology raises concerns about representativeness, particularly due to the mix of online and face-to-face data collection.

Response: Thank you for your insightful comment. We have detailed the process of sampling and data collection in the Methodology section (lines 114 - 145). To minimize the bias of under- and overrepresentation, we have utilized the weight adjustment technique (lines 232 - 241).

Additionally, some statistical interpretations require further clarification, and the discussion on policy implications could be expanded to provide actionable recommendations. Addressing these concerns will enhance the overall impact and credibility of the study.

Response: Thank you for your insightful comment. We have carefully interpreted the findings of our study (lines 259 - 273) and provided actionable recommendations in the Conclusion section (lines 411 - 429).

Specific Comments

Major Comments

1. The study employs both online and face-to-face data collection. However, the online survey may have

overrepresented educated and tech-savvy individuals, while the face-to-face survey followed quota sampling.

Response: Thank you for your feedback. However, the digital divide in the country was considered when finalizing the methodology, and the ratio for face-to-face and online surveys was kept at 2:1, taking this into account. This is detailed in the Methodology section (lines 89 - 90 and 98 - 118).

2. Clarify the adjusted odds ratio (aOR) interpretation. Some aOR values are close to 1, making practical significance questionable.

Response: Thank you for your feedback. We have reanalyzed the data with weighted sample size to address your comment (Table 2).

3. The impact of administrative divisions (eg, Sylhet having 4× higher WTP) should be further discussed. Are these differences due to economic, cultural, or policy variations?

Response: Thank you for your valuable comment. In the discussion section, we further discuss the impact of administrative divisions (lines 355 - 368).

4. While the study suggests subsidized vaccination programs, it would be helpful to compare findings with other low- and middle-income countries' WTP trends.

Response: Several studies have been added to the discussion of low- and middle-income countries' WTP trends (lines 318 - 328).

Minor Comments

5. Ensure table captions clearly describe what is presented (eg, Table 2 should explicitly state that it presents logistic regression results).

Response: Thank you for your suggestion. We have made changes to the table captions to reflect what they present.

6. Some sections contain grammatical errors and awkward phrasing (eg, "knowledge about the vaccine, vaccine process, conspiracy beliefs, behavioral practice, attitude toward a vaccine"; this list is repetitive and unclear).

Response: Thank you for your insightful comment. We have revised the manuscript to minimize grammatical errors and improve awkward phrasing.

Reviewer BN [3]

This paper addresses an important and timely topic—WTP for COVID-19 vaccines in a developing country context. Understanding WTP is essential not only for informing current vaccine financing strategies but also for shaping policies related to equitable vaccine access in response to future public health challenges. The study is well-conceived and provides valuable insights into vaccine affordability and public perception in Bangladesh. With some refinements in presentation, statistical interpretation, and policy framing, the paper will be well-positioned for publication.

The abstract would benefit from being more concise and should more clearly highlight the key policy implications of the findings. Additionally, the statistical interpretation of the aORs requires careful attention. Several aORs are reported with values close to 1 (eg, family income aOR 1.0, P=.039; vaccine knowledge aOR 1.1, P=.003; behavioral practices aOR 1.1, P<.001), suggesting minimal effect sizes, yet they are statistically significant. While such significance may be driven by the large sample size, reporting CIs would allow for a more meaningful interpretation of the strength and direction of these associations.

Response: Thank you for your valuable comment. We've rewritten the abstract to address your comment. We have also reanalyzed the data using a weighted sample size, which addresses the issue related to the aORs (Table 2) and reports CIs for more meaningful interpretation.

The paper would also benefit from greater clarity around the construction of variables and the underlying measurement models. It is unclear how multiple survey items were combined to form factors such as knowledge, attitudes, and behavioral constructs. Using exploratory factor analysis could be beneficial to validate the grouping of items into coherent factors and strengthen construct validity. Providing factor loadings or at least a brief description of the item-grouping process would enhance the methodological transparency of the study.

Response: Thank you for your insightful comment. However, please note that the objective of this paper is not to develop a scale on these issues. Instead, we aimed to identify the correlates of WTP. If we focus on the exploratory factor analysis and confirmatory factor analysis for these scales, the readers may get distracted. Thus, we focused only on the reliability analysis of the items used to measure these scales using Cronbach α, which has been discussed in the Methodology section (lines 123 - 186).

Another area for improvement involves the reporting of the income variable. In both Table 1 and Table 2, income appears to be modeled as a continuous variable, but the unit of measurement is not specified. Without this information, it is difficult to interpret an odds ratio of 1.0 meaningfully. If income is measured in small units (eg, Bangladeshi taka), the impact of each unit increase would be negligible. Categorizing income into meaningful brackets (eg, low, middle, high) and using those categories in logistic regression would make the results more interpretable and policy relevant.

Response: Thank you for your comment. We've categorized the income variable, and the result now appears more interpretable and policy relevant (Tables 1 and 2).

Additionally, the CIs for some variables in Table 2—such as income and COVID-19 vaccine conspiracy beliefs—appear to suggest nonsignificance, yet they are reported as significant. This inconsistency should be carefully reviewed and clarified.

Response: Thank you for your valuable comment. We have categorized the income variable and revised the regression models to solve these issues (Table 2).

Some of the measured constructs, such as knowledge and perceived susceptibility, show relatively low internal consistency (eg, Cronbach α of 0.612 and 0.657, respectively). It would be helpful for the authors to explain why these values are

considered acceptable in this context or to discuss efforts made to improve reliability through item refinement or scale revision.

Response: Thank you for your valuable comment. We've explained the issues of accepting low internal consistency in the Methodology section (lines 148 - 153).

Furthermore, the combination of nonprobability online sampling and quota sampling should be more clearly justified. While practical during a pandemic, it raises concerns about representativeness and potential sampling bias, which should be acknowledged more explicitly in the Discussion.

Response: Thank you for your valuable feedback. We acknowledge the raised concern regarding the combination of nonprobability online sampling and face-to-face sampling. Our study employed this hybrid sampling method to ensure adherence to safety protocols amid the COVID-19 pandemic. Through an online survey, we quickly reached a large audience. However, later, we conducted face-to-face data collection using sampling criteria that ensured the representativeness of the sample, thereby determining the population's national representation in terms of age, sex, residence, division, and marital status.

The manuscript would also benefit from a thorough review for minor language and formatting issues. For instance, the phrase "explains explains" on page 13 should be corrected. Variable labels and descriptions in tables should be presented clearly and consistently.

Response: Thank you for your comment. We have reviewed the manuscript and made the necessary corrections accordingly.

Reviewer BM [4]

1. In lines 79 and 80 of the manuscript [1], it is confusing why this wouldn't be considered nationally representative if the data collection was conducted online.

Response: Thanks for pointing this out. We were discussing the limitations of existing studies. We have now described the issues more carefully (lines 77 - 81).

2. As around 50% of the people are not interested in paying for the vaccine, this result should be considered with caution.

Response: Thank you for your insightful comment. We acknowledge that WTP for a vaccine is context dependent. Our study's results may be influenced by unique sociodemographic and cultural dynamics that appeared during data collection. We have mentioned this as a limitation in the Discussion section.

References

1. Vij J. Peer review of "Willingness to Pay for the COVID-19 Vaccine and Its Correlates in Bangladesh: Cross-Sectional Study". JMIRx Med 2025;6:e79353. [doi: 10.2196/79353]
2. Hossain MB, Alam MZ, Islam MS, et al. Willingness to pay for the COVID-19 vaccine and its correlates in Bangladesh: cross-sectional study. JMIRx Med 2025;6:e69827. [doi: 10.2196/69827]
3. Kabir E. Peer review of "Willingness to Pay for the COVID-19 Vaccine and Its Correlates in Bangladesh: Cross-Sectional Study". JMIRx Med 2025;6:e79355. [doi: 10.2196/79355]
4. Hoque E. Peer review of "Willingness to Pay for the COVID-19 Vaccine and Its Correlates in Bangladesh: Cross-Sectional Study". JMIRx Med 2025;6:e79354. [doi: 10.2196/79354]

Abbreviations

aOR: adjusted odds ratio
WTP: willingness to pay

Edited by F Wu; submitted 19.06.25; this is a non–peer-reviewed article;accepted 19.06.25; published 15.08.25.

Please cite as:
Hossain MB, Alam MZ, Islam MS, Sultan S, Faysal MM, Rima S, Hossain MA, Mamun AA, Mamun AA
Authors' Response to Peer Reviews of "Willingness to Pay for the COVID-19 Vaccine and Its Correlates in Bangladesh: Cross-Sectional Study"
JMIRx Med 2025;6:e79352
URL: https://xmed.jmir.org/2025/1/e79352
doi:10.2196/79352

Author's Response to Peer Reviews of "Effect of Thermal and Vibration Changes on Automated External Defibrillator Circuit Boards: Finite Element Analysis Study"

Saidi Olayinka Olalere, MSc

Department of Mechanical Engineering, Georgia Southern University, 1332 Southern Drive, Statesboro, GA, United States

Corresponding Author:
Saidi Olayinka Olalere, MSc
Department of Mechanical Engineering, Georgia Southern University, 1332 Southern Drive, Statesboro, GA, United States

Related Articles:

Companion article: https://arxiv.org/abs/2309.14747v1

Companion article: https://med.jmirx.org/2025/1/e80142

Companion article: https://med.jmirx.org/2025/1/e80137

Companion article: https://med.jmirx.org/2025/1/e53208

Abstract

(JMIRx Med 2025;6:e80135) doi:10.2196/80135

KEYWORDS

circuit board; automated external defibrillator; heart; cardiology; vibration; thermal changes; medical devices

This is the authors' response to peer-review reports for "Effect of Thermal and Vibration Changes on Automated External Defibrillator Circuit Boards: Finite Element Analysis Study."

Round 1 Review

Reviewer R [1]

General Comments

This paper [2] considered the vibration and thermal analysis of a modeled circuit board of an automated external defibrillator (AED) using Ansys. The vibration failure in the modeled circuit board with four rigid supports was pronounced starting at the taller components, including the capacitor. The board was reinforced to an 8–rigid support system, reducing the failure around the rigid supports. The thermal failure started from the battery position, causing thermal dissipation to other parts of the board, ultimately leading to the failure of the circuit board and the AED.

Specific Comments

Major Issues

1. "The goal is to analyse the effect of vibration and thermal experience on the AED based on its operation." Is this your research statement? I suppose the topic suggests you analyzed the modeled circuit board, not the overall AED system. If not, how did you measure the overall effect on the AED?

Response: The research was a simulation of how a modeled circuit board of AED is affected by vibration and thermal conditions. This was the basis of the research, and the analysis was performed on these premises.

2. The author may also need to discuss the importance of the circuit boards in an AED in the Introduction.

Response: The AED circuit board for the analysis will be a material made of epoxy FR-4 with a length of 254 mm and width of 216 mm, while the thickness is 0.5 mm. The components of the circuit board include a capacitor, microcontroller, flash memory, analogue digital converter, field programmable gate array, processor, audio controller, inductor, and more.

3. The figures will need a little more discussion.

Response: The figures were explained individually. Also, a Discussion heading was created to discuss the findings before a conclusion was used to end the paper.

Minor Comments

4. "Vibration and Thermal Analysis on Modeled Circuit Board of Automated External Defibrillator (AED) Medical Device" will likely communicate the title better.

5. I find it a bit difficult to understand this line: "Fatigue failure under sinusoidal vibration loading for component by comparing the vibration failure test, FEA, and theoretical test (Y.S.Chen, 2008)." What did you want to say?

Response: This is a research reference to Chen et al [3], who explained that failure was experienced from vibration loading on components when the simulation and experimental testing were performed.

6. Figure 1 will need relabeling. The labeling seems to cover some parts of the board. A transparent background could help.

Response: The majority of the labels are outside of the board, and a transparent background could distort the labeling and look like wording on the board.

Anonymous [4]

General Comments

I have read the submitted manuscript to your journal entitled "Analysis of Vibration and Thermal of a Modeled Circuit Board of Automated External Defibrillator (AED) Medical Device." The author utilized finite element analysis to simulate static and dynamic testings in determining the vibration and thermal effects on the operations of the AED medical device.

Based on the outcomes of this conducted study and the potential benefits from this study, I would recommend this manuscript to be published in your reputable journal after the minor comments are properly addressed.

Specific Comments

Minor Comments

There are 4-member and 8-member support. The author should provide more evidence on how these affect the analysis results.

Response: A physical AED was used as a prototype to model the circuit board. A physical AED is always a 4-member support, which was the basis of the research. For further understanding, an 8-member support was modeled to understand any change with respect to the vibration with a change in the support.

This manuscript will benefit from other works on the failure modes of materials. The author is encouraged to expand the Literature Review section and also provide more references.

Response: The literature has been expanded and necessary references included.

The grammar should be refined. Some of the grammar in this manuscript should be corrected. For example, "This study was performed to analysis the vibration..." "Analyse." Fig 14 = Fig. 14 or Figure 14. The unit of temperature is degree c; the c is capitalized.

Response: The grammatical errors in the abstract and figures have been corrected.

When citing a research paper within the manuscript, it is the first/lead author who should be named in the "Name et al" format; the author should use a consistent citation format.

Response: The citation has been corrected.

The author should remove parentheses from the manuscript title: "Analysis of Vibration and Thermal of a Modeled Circuit Board of Automated External Defibrillator Medical Device."

Response: Parentheses have been removed from the title.

For reproducibility by other researchers, the author should consider providing simulation data as supplementary information.

Response: For reproducibility, the materials and methods provide the necessary information methodology and requirements in terms of model design, mesh selection, variables and values, and more.

The author should properly cite other works where applicable. For example, "From other research results, it can be verified that the natural frequencies..." The author needs to insert appropriate citations for comments like these.

Response: The necessary citation has been included as recommended.

Round 2 Review

Reviewer R

1. The recommendation about explaining how the author measured the overall effect of the analysis on the AED or the board and the recommendation that the author should provide more evidence on how the 4-member and 8-member supports affect the analysis result have not been answered or addressed in the manuscript. The author should consider these.

Response: The 4 and 8 members were explained more with tables and figures, which can be found in the manuscript. The 8 members were explicated in Figures 12 and 13.

2. The author will also need to be consistent. Is Figure 5 the same as Figure 5 or Fig. 5? It should be corrected for all other instances.

Response: This is a generally acceptable concept in manuscripts, where Figure (in full) is used in the naming convention, while Fig. is used in further explanation.

3. Additional citations might be needed in the work; it still looks like over 40% of the citations are 15 years or older. Also, "et al.," should be in italics with a period after the "al." The Discussion also seems not well discussed in relation to previous works.

4. The template format should also be considered carefully.

Response: Only related citations were included in the manuscript. This manuscript shows a blend of early research and recent research that aligned with the purpose of the paper. I have 10 references that are more detailed toward the research.

References

1. Akinfenwa A. Peer review of “Effect of Thermal and Vibration Changes on Automated External Defibrillator Circuit Boards: Finite Element Analysis Study”. JMIRx Med 2025;6:e80142. [doi: 10.2196/80142]
2. Olalere SO. Effect of Thermal and Vibration Changes on Automated External Defibrillator Circuit Boards: Finite Element Analysis Study. JMIRx Med 2025;6:e53208. [doi: 10.2196/53208]
3. Chen YS, Wang CS, Yang YJ. Combining vibration test with finite element analysis for the fatigue life estimation of PBGA components. Microelectronics Reliability 2008 Apr;48(4):638-644. [doi: 10.1016/j.microrel.2007.11.006]
4. Anonymous. Peer review of “Effect of Thermal and Vibration Changes on Automated External Defibrillator Circuit Boards: Finite Element Analysis Study”. JMIRx Med 2025;6:e80137. [doi: 10.2196/80137]

Abbreviations

AED: automated external defibrillator

Edited by T Leung; submitted 04.07.25; this is a non–peer-reviewed article;accepted 04.07.25; published 19.08.25.

Please cite as:
Olalere SO
Author’s Response to Peer Reviews of “Effect of Thermal and Vibration Changes on Automated External Defibrillator Circuit Boards: Finite Element Analysis Study”
JMIRx Med 2025;6:e80135
URL: https://xmed.jmir.org/2025/1/e80135
doi:10.2196/80135

Authors' Response to Peer Reviews of "Rapidly Benchmarking Large Language Models for Diagnosing Comorbid Patients: Comparative Study Leveraging the LLM-as-a-Judge Method"

Peter Sarvari, MEng, MS, MBA; Zaid Al-fagih, BSc, MBBS, MPP

Rhazes AI, First Floor, 85 Great Portland Street, London, United Kingdom

Corresponding Author:
Peter Sarvari, MEng, MS, MBA
Rhazes AI, First Floor, 85 Great Portland Street, London, United Kingdom

Related Articles:

Companion article: https://www.preprints.org/manuscript/202409.0688/v3

Companion article: https://med.jmirx.org/2024/1/e69830

Companion article: https://med.jmirx.org/2025/1/e67661

(JMIRx Med 2025;6:e81235) doi:10.2196/81235

KEYWORDS

LLM; GPT-4; Gemini; Claude; RAG; clinical medicine; diagnosis; diagnostic ability of LLMs; large language model; AI in medicine; AI in healthcare; retrieval-augmented generation; artificial intelligence

This is the authors' response to the peer-review report for "Rapidly Benchmarking Large Language Models for Diagnosing Comorbid Patients: Comparative Study Leveraging the LLM-as-a-Judge Method."

Many thanks for your thoughtful comments [1] on our submission [2]. We spent time to carefully rewrite the article so it addresses the editorial comments and those from the live review.

Major Concerns and Suggested Improvements

Title Revision

The current title does not fully capture the scope of the study; hence, it needs to be reconsidered. It would be nice if abbreviations were avoided in the title. Therefore, we recommend the authors of the study change "LLM" to "Large Language Models" in the title.

Response: Title has been revised.

Abstract and Introduction Clarity

The abstract and the introduction lack a clear statement of the study's aim. It is therefore expedient to revise the abstract to include the objectives, methodology, key results, and conclusion. The Introduction should have a clear research aim.

Note: During the call, the authors shared that a revised version of the abstract had been generated and shared so it is possible that the latest version has addressed this concern. We invite the authors to share their updated version in the comment section of this review.

Response: The abstract and introduction have been significantly rewritten since the live review.

Physician Comparison With Large Language Models

The study does not explore how diagnoses differ from physicians using only the data provided to the large language model (LLM). It would be advisable to include a comparative analysis to evaluate diagnosis accuracy and the prioritization of additional tests between physicians and LLMs.

Furthermore, the absence of actual data around patient history and other diagnostic parameters beyond what was reported in billing reports (reported as "ground truth" in the study) is a weakness. This can lead to an incomplete or partial diagnosis being labeled as the final diagnosis, leading to miscalculations about the accuracy of LLMs.

Response: We added this as a limitation of the study.

Model Selection Rationale and Evaluation Metrics

The Methods section is limited in its description of the methodology used in the study. It would be helpful to include more information on the rationale for the model selection and describe differences between GPT-4 variants to help readers understand the comparative approach.

Furthermore, the choice of "hit rate" as the primary evaluation metric is unclear, and its limitations are not discussed in sufficient detail. It would be helpful if the choice of hit rate over other metrics (eg, precision or F1-score) as well as the

limitations the hit rate may introduce were discussed more thoroughly.

Response: We expanded the Methods section to include more details about the automated evaluation, retrieval-augmented generation (RAG), and the model versions used. The rationale for choosing hit rate is now described in detail in the methods.

Methodology and RAG Integration Details

The role of RAG in the diagnostic process, including how relevant information was retrieved and implemented to enhance the diagnostic process and performance needs to be elaborated further as it constitutes a novel part of the study. This issue was highlighted as one of the particular concerns, as without more details, many questions remain unanswered and that could compromise the credibility of the study.

Response: We added more details on our RAG methodology, simplified and reran all experiments, and confirmed their statistical significance. The exact sections retrieved (10 out of 32 chunks) vary between the 1000 patients and experiment runs.

Data Interpretation and Population-Specific Reference Ranges

Reference ranges used for diagnoses are not adequately explained. The authors are encouraged to clarify if the reference ranges are population-specific or if they align with the dataset characteristics.

In general, reviewers suggest authors add more details about the nature of the data beyond referring to them as "test results" in the manuscript. For example, it would be helpful to know more about the meaning and interpretation of the homogeneity of the test results and the implications of it on the evaluation of the method.

Provide a statistical analysis to demonstrate that the differences in diagnostic hit rates for the LLMs are statistically significant in the range of 98.5-99.8.

Response: We use the most recent American Board of Internal Medicine laboratory reference ranges as referenced. Please see [3].

Discussion

It would be helpful to discuss why GPT-4.0 and Claude 3.5 Sonnet performed better than others, potentially due to architectural differences or data training sources.

It would also be important to discuss why specific diagnoses (eg, diabetes) were among the best hits and most frequent misses.

Response: Unfortunately, we do not have access to the exact architecture or training data of closed-source models provided by companies like OpenAI or Anthropic. The exact reasons the diagnostic models gave for hits and the assessor models gave for missed diagnoses are available from the GitHub repository directly. All results are saved as a CSV under the "data" folder. If the editors would like, we can give an example about how one would go about analyzing these in the appendix.

Limitations of Study Design

The limitations could be explicitly outlined in a separate section of the Discussion for transparency and clarity. For example, the authors may include a discussion around the fact that the sample size (1000 patients) may be too small to generalize the findings, potential issues related to relying on billing reports as ground truth, and considerations of hallucinations or failure scenarios of LLMs in real-world settings. In such a section, the authors may also explore ideas related to using larger and more diverse datasets in similar future research.

Response: The Limitations section has been greatly expanded.

Figures and Tables

The figures and tables in the study lack clarity and, at times, key information (eg, patient demographics, disease types are missing). The authors are advised to add clarity to the data visualizations, label axes, and include interpretive analyses to all figures, but in particular for Tables 1 and 2, and Figure 2. They are also advised to discuss specific trends such as frequent misses for certain conditions.

Response: Tables and figures have been changed significantly since this comment was made.

Reproducibility

The reproducibility of the study is hindered by the lack of clear documentation on LLM settings, dataset transformations, and code. It is suggested that the authors provide the full details of the LLM configurations, processing steps, and code availability.

For example, it would be helpful to know the rationale for limiting the LLM output tokens to 4096. How could this be relevant to the "human" diagnostic process? Were some predictions judged "more likely" than others?

Response: Everything needed for reproducibility has been shared on the GitHub repository. If the editors would like, we can give an example about how one would go about running a new experiment in the appendix.

Bias and Real-World Application

Potential biases in LLM predictions and challenges in clinical adoption are not addressed in the study. It is advised that the authors add a section on potential biases and practical integration challenges. They need to include future work on improving model robustness and fairness.

Response: Thank you—it was added to the Limitations section.

Minor Concerns and Suggested Improvements

Abbreviation Usage

Key abbreviations (eg, LLM, RAG) were not defined at first use. The authors are encouraged to define all abbreviations when first indicated in the abstract and body of the study (eg, "electronic health records (EHRs)" when first mentioned, then "EHR" at later mentions).

Response: Thank you—RAG and EHR have been defined.

Language

There are several typos and some grammatical errors, incomplete sentences, and contractions that reduce the readability of the study; hence, the authors are encouraged to consider thorough proofreading and editing to improve the reading experience and interpretation of the study. This is a minor concern that may be well addressed by the copyeditors of the journal that will publish the manuscript.

Response: Thank you—we made further edits to make the manuscript more readable.

Ethical Statement Clarity

The ethical considerations for using MIMIC-IV data are not explicitly referenced. The authors should state that the dataset is deidentified and describe access restrictions for researchers. Some reviewers had concerns about the need for ethical approval given the use of patient data, but others reported that ethical approval may not be needed given the public nature of the data used.

Furthermore, it would be helpful to add a discussion around the potential risk of bias introduced by LLMs and its large implications on diagnosis and the field of medicine at large.

Response: Added under Data Availability statement.

False-Positive and False-Negative Rates

The explanation of false-positive and false-negative rates in the study is inadequate; hence, the authors are invited to include specific examples and explanations of why certain diagnoses were misclassified.

Response: A specific example was given in our previous study [4]. Could you please clarify which part of the explanation was inadequate?

Conclusions

The author should consider adding a section that examines potential biases in LLM predictions and the practical challenges of using these models in hospital settings. Furthermore, it would be helpful to further highlight practical takeaways or future directions, emphasizing actionable insights and specific areas for future research (eg, integrating multimodal data sources or fine-tuning models with diverse clinically annotated datasets).

Response: While we did not add a separate section about biases, we direct the reader to a comprehensive review on this topic. Future directions about hospital implementation and improving the limitations have been addressed.

Comparative Model Performance

Performance differences between models in the study are not sufficiently elaborated on in the Discussion. The authors are invited to explore why certain models performed better, considering architectural differences and training data sources. Reviewers also advised authors to consider human vetting for the evaluation to provide an additional layer of confidence to get the experts to reflect on the LLM answers and explanations.

Response: Unfortunately, we do not have access to the exact architecture or training data of closed-source models provided by companies like OpenAI or Anthropic.

Hyperbolic Language

Words like "stunning" are overly subjective. The use of neutral language is advised in the manuscript, and the authors are invited to justify claims with supporting data.

Response: Thank you—they have been removed.

Dataset Limitations

Rare diseases may not be adequately represented in the study. The authors should address how dataset limitations affect diagnostic performance and include rare disease cases in future studies.

Response: The Limitations section has been greatly expanded.

Citations to Methods and Tools

Where possible, add citations to specific LLM and RAG tools used, such as technical references from Google, OpenAI, etc, to aid readers in finding more information on these tools.

Authors are advised to complete their statements instead of just including a citation. For example, "In this case, the further tests the LLM is instructed to suggest [2] are of crucial importance to understand exact disease pathology."

Provide an explanation of the sentence "NEJM Case Challenges are notoriously hard" and provide a reference. Potentially, reconsider the use of the extreme adverb "notoriously"—perhaps "well known to be."

Response: Statements have been completed. We are happy to add references to all models mentioned in the article if the editors agree that this would enhance the quality of the paper. Most model documentations can be found simply by googling the model name and version, both of which we have provided. Extreme words have been removed according to our best judgment.

Presentation of Methods

For readability, reformat the list of LLMs used into a table with separate columns for name, version, and settings.

Response: We are not sure that this would enhance the flow of the manuscript. If the editors agree, however, we are happy to make such a modification. The settings are largely synonymous across models; it is just the model names and versions that differ.

The authors would like to thank all peer reviewers for their thoughtful feedback and great contribution toward making this manuscript better.

References

1. Saderi D, Mahmoud RSG, Bender G, et al. Peer review of "Towards Evaluating the Diagnostic Ability of LLMs (Preprint)". JMIRx Med 2024;5:e69830. [doi: 10.2196/69830]

2. Sarvari P, Al-fagih Z. Rapidly benchmarking large language models for diagnosing comorbid patients: comparative study leveraging the LLM-as-a-judge method. JMIRx Med 2025;6:e67661. [doi: 10.2196/67661]
3. ABIM laboratory test reference ranges. American Board of Internal Medicine. 2025 Feb. URL: https://www.abim.org/Media/bfijryql/laboratory-reference-ranges.pdf [accessed 2025-08-13]
4. Sarvari P, Al-Fagih Z, Ghuwel A, Al-Fagih O. A systematic evaluation of the performance of GPT-4 and PaLM2 to diagnose comorbidities in MIMIC-IV patients. Health Care Sci 2024 Feb;3(1):3-18. [doi: 10.1002/hcs2.79] [Medline: 38939167]

Abbreviations

EHR: electronic health record
LLM: large language model
RAG: retrieval-augmented generation

Edited by A Schwartz; submitted 24.07.25; this is a non–peer-reviewed article;accepted 24.07.25; published 29.08.25.

Please cite as:
Sarvari P, Al-fagih Z
Authors' Response to Peer Reviews of "Rapidly Benchmarking Large Language Models for Diagnosing Comorbid Patients: Comparative Study Leveraging the LLM-as-a-Judge Method"
JMIRx Med 2025;6:e81235
URL: https://xmed.jmir.org/2025/1/e81235
doi:10.2196/81235

Authors' Response to Peer Reviews of "Challenges in Implementing a Mobile AI Chatbot Intervention for Depression Among Youth on Psychiatric Waiting Lists: Randomized Controlled Study Termination Report"

Junichi Fujita[1], PhD, MD; Yuichiro Yano[2], PhD, MD; Satoru Shinoda[3], PhD; Noriko Sho[4], PhD, MD; Masaki Otsuki[5], MD; Akira Suda[2], PhD, MD; Mizuho Takayama[1], MHSW, BSc; Tomoko Moroga[1], RN, BSc; Hiroyuki Yamaguchi[6], PhD, MD; Mio Ishii[6], PhD, MD; Tomoyuki Miyazaki[7], PhD, MD

[1]Department of Child Psychiatry, Yokohama City University Hospital, 3-9, Fukuura, Kanazawa-ku, Yokohama, Japan

[2]Psychiatric Center, Yokohama City University Medical Center, Yokohama, Japan

[3]Department of Biostatistics, Yokohama City University School of Medicine, Yokohama, Japan

[4]Kanagawa Children's Medical Center, Yokohama, Japan

[5]Fujisawa City Hospital, Fujisawa, Japan

[6]Department of Psychiatry, Yokohama City University School of Medicine, Yokohama, Japan

[7]Center for Promotion of Research and Industry-Academic Collaboration, Yokohama City University, Yokohama, Japan

Corresponding Author:
Junichi Fujita, PhD, MD
Department of Child Psychiatry, Yokohama City University Hospital, 3-9, Fukuura, Kanazawa-ku, Yokohama, Japan

Related Articles:

Companion article: https://www.medrxiv.org/content/10.1101/2024.11.25.24317880v1

Companion article: https://med.jmirx.org/2025/1/e82071

Companion article: https://med.jmirx.org/2025/1/e82073

Companion article: https://med.jmirx.org/2025/1/e82074

Companion article: https://med.jmirx.org/2025/1/e70960

(JMIRx Med 2025;6:e82083) doi:10.2196/82083

KEYWORDS

randomized controlled trial; AI chatbot; acceptance and commitment therapy; mental health; psychiatry; children; adolescents; Japan

This is the authors' response to peer-review reports for "Challenges in Implementing a Mobile AI Chatbot Intervention for Depression Among Youth on Psychiatric Waiting Lists: Randomized Controlled Study Termination Report."

Round 1 Review

Reviewer E [1]

Thank you for inviting me to review this paper [2]. However, my suggestion would be that this paper should be rejected. I am very cognizant of publication biases, and I am a firm believer that the publication of negative results is very important. I therefore have no problem with the fact that the sample decreased to zero. However, I do believe that more detail is needed in terms of why people disengaged. The rationale for the paper is set up as efficacy of the intervention, but the main message of the paper is that the sample declined. I would therefore like more emphasis on qualitative interviews that examined why people disengaged. Follow-up work such as this would make a very interesting paper.

Response: We appreciate this insightful comment. While our study did not initially include structured qualitative interviews, we recognize that a more systematic approach to understanding disengagement barriers would provide deeper insights. In response, we have revised the Strengths and Limitations section in the Discussion to explicitly acknowledge this limitation and propose the incorporation of structured exit interviews or surveys in future studies. Additionally, we have expanded the Results section to include details of participant withdrawal,

particularly emphasizing parental reports of the adolescent's distress regarding online participation.

Reviewer I [3]

General Comments

This paper describes the results of a parallel group randomized controlled trial that examined the feasibility of an artificial intelligence (AI) chatbot-led mental health intervention to support pediatric patients on the psychiatry waitlists in Japan. The article is well-written and organized, and the objectives of the study are clearly stated. Methodology elements such as eligibility criteria, information sources, and data collection process are clear. A clear list of outcomes and variables for which data were researched is presented. The authors provide an important contribution to the field by reporting on factors that challenge adolescents' engagement in digital mental health interventions and providing meaningful recommendations for future research.

Specific Comments

Major Comments

1. How many chatbots were shortlisted, and why was emol favored over the others, given the selection criteria? (Under AI Chatbot Selection Process.)

Response: We have clarified the AI Chatbot Selection Process section, detailing that multiple AI chatbots were reviewed based on predefined selection criteria. AI chatbot emol was chosen due to its integration of acceptance and commitment therapy (ACT) principles, user engagement features, and prior application in mental health settings.

2. How are the six core processes of ACT delivered in the AI chatbot (under Intervention Group)? Expand more on each section. How does the session meet the core processes of ACT—acceptance, cognitive defusion, being present, self as context, values, and committed action?

Response: We have expanded the Intervention Group section to describe how each ACT process—acceptance, cognitive defusion, being present, self-as-context, values, and committed action—is incorporated into specific chatbot sessions. Additionally, we have created a supplementary table that provides a comprehensive overview of the session structure, showing how each session aligns with specific ACT processes. The table details the content types (videos, comics, written practices) used to deliver these therapeutic concepts in an engaging and accessible format for adolescents.

3. How was the section structured? Did adolescents go through modules? Could they write anything to the chatbot, or was the content predefined? Were the sessions sequentially delivered or not? Could they access previously completed modules or track their progress?

Response: The revised Intervention Group section now clarifies session progression, user interaction (predefined vs open-ended responses), and module accessibility. We have also created a supplementary figure illustrating the actual interface of the emol application, including screenshots of conversations between the AI character Roku and users. This figure demonstrates how therapeutic concepts are introduced in a conversational age-appropriate manner, and how users engage with the app through both structured exercises and dialogues.

4. Were there any safeguarding links and referral contacts built into the chatbot in case participants needed additional support beyond those offered by the chatbot? If yes, I recommend including it under the ethics paragraph.

Response: We have added details in the Ethical Considerations section, confirming that emergency support information was provided to all participants and that the research team had a protocol for directing participants to appropriate psychiatric services if necessary.

5. How were you planning to investigate engagement? Would you report on the frequency of use, number of interactions with the chatbot, or amount of content visualized by participants? Even though the study's main questions are not focused on engagement, I suggest that the authors consider including an engagement outcome paragraph right after the secondary outcomes.

Response: The Data Management section now specifies that engagement was tracked through AI chatbot emol's data logs, recording total usage time, average daily usage, session progression, and last completed session. These engagement metrics were intended to assess both usage patterns and their relationship with depressive symptom changes.

Minor Comments

6. I recommend moving all hyperlinks to the appendix and including an image of the chatbot. I also recommend that authors include an image of the intervention delivered through the hospital website.

Response: We have relocated hyperlinks to the appendix and created a supplementary figure showing the chatbot interface and interaction examples. The figure includes screenshots of the AI character Roku and demonstrates key features of the app, including how it introduces ACT concepts, guides users through exercises, and provides supportive feedback. This visual representation helps clarify the user experience and the app's design elements specifically tailored for adolescents.

7. Please state the statistical methods used to deal with missing data.

Response: The Statistical Analysis section now explicitly states the methods used to handle missing data. Missing data were analyzed as observed without imputation, with the primary analysis performed on the full analysis set. Given the exploratory nature of this study and the lack of prior research in this specific population, no imputation was conducted, ensuring that the results accurately reflect the available data without introducing assumptions through imputation methods.

8. In the Discussion, you argue that young people prefer online mental health support over in-person support [4]. I believe you could discuss this a bit more in your Introduction paragraph to strengthen your discussion regarding the potential gap online services could fill.

Response: The Introduction section has been revised, particularly in the first paragraph, to incorporate a discussion on the preference for online mental health support among youth. This revision, supported by existing literature, emphasizes the increasing demand for accessible and scalable digital mental health interventions.

9. I recommend including a paragraph under the Introduction on previous Japanese studies focusing on chatbot-led or digital mental/public health interventions to provide an overview of the current population uptake of digital health interventions.

Response: A paragraph summarizing prior chatbot-led interventions in Japan has been added to the Introduction section, specifically in the fifth paragraph. This revision also incorporates cultural factors unique to Japan, providing a more comprehensive understanding of how chatbot interventions are perceived and utilized in this context.

Reviewer M [5]

General Comments

The topic and objectives of the study are certainly interesting, as depression among young individuals is an increasingly pervasive and growing problem globally, exacerbated by the COVID-19 pandemic, as the authors themselves point out. Furthermore, the use of AI to support traditional methods of treating this condition makes the study topical. The paper is well-written and comprehensible throughout; the supporting bibliography is adequate; it has a good methodological approach, with clear and well-defined objectives, and an accurate description of the inclusion and exclusion criteria for participants. Although the statistical analyses planned by the authors are consistent with the objectives they have defined, the lack of availability of data on which to carry out these analyses and, therefore, the absence of results does not allow an evaluation of this specific aspect. However, the authors have posited potential explanations for instances of nonadherence to the intervention protocol, which are substantiated by extant literature on the subject, therefore apprising the reader of the possible limitations of this type of intervention in this specific population that fulfills certain inclusion criteria. The paper thus provides a cue and guidance for future studies in this field. Lastly, as stated in the major comments below, the major shortcoming of this study is the lack of clarity as to whether the authors used an active or nonactive control group.

Specific Comments

Major Comments

1. In the Study Design paragraph, the authors stated that the control group would receive standard care (making it an active control group), while in the Control Group paragraph, they stated that they would receive general mental health information and would undergo online evaluations and diary recordings (making it a nonactive control group). It is not clear if the authors deem these two procedures similar. In the event that they do not regard them as analogous, it would be beneficial to ascertain which of the two would have been delivered to the control group. Furthermore, it would be appreciated if the authors could provide an explanation and make the appropriate adjustments in the manuscript about (1) what standard care would have comprised and (2) what is the nature of the short video programs that participants received as general mental health information, in order to enable the reader to ascertain whether they are informational videos, mental health support videos, etc.

Response: We have revised the Study Design and Control Group sections to clarify that the control group received general mental health information via a publicly available website, not standard psychiatric care. The educational materials include child-friendly videos featuring a psychiatrist explaining mental health topics using animated characters.

Additionally, the Introduction (fifth paragraph) now provides a more detailed discussion of prior chatbot-led interventions in Japan. The Methods section has been expanded to specify the content of the videos and the online evaluations used in the control condition, including voice analysis and writing pressure measurements, ensuring transparency in the study's evaluation process.

Round 2 Review

Reviewer M

General Comments

I would like to express my gratitude to the authors for implementing the requested revisions, which have served to enhance the clarity and thoroughness of the manuscript. Still, there are some elements that, in my view, would benefit from modification.

Specific Comments

Major Comments

1. Supplementary Table 1 and the supplementary figure are missing.

Response: We apologize for the oversight. We have now uploaded Supplementary Table 1 and the supplementary figure as separate multimedia appendix files and referenced them in the manuscript as "Multimedia Appendix 1" and "Multimedia Appendix 2," respectively.

2. The sentence "AI chatbot emol features a friendly character name 'Roku'" is redundant, as the same concept is repeated in the preceding sentence (in the AI Chatbot Selection Process paragraph).

Response: Thank you for pointing this out. We have removed the redundant sentence "AI chatbot emol features a friendly character named 'Roku,' who guides users through ACT-based conversations in a relatable manner." to improve conciseness.

Before: "AI chatbot emol's design prioritizes accessibility and engagement, particularly for young users, by featuring a friendly AI character named 'Roku.' AI chatbot emol features a friendly character named 'Roku,' who guides users through ACT-based conversations in a relatable manner."

After: “AI chatbot emol’s design prioritizes accessibility and engagement, particularly for young users, by featuring a friendly AI character named ‘Roku.’”

3. The following sentence is repeated twice: “Weekly online assessments were conducted at Week 0, during the intervention period, and at Week 9” (in the Intervention Group paragraph).

Response: Thank you for noting the repetition. We have removed the duplicated sentence to streamline the narrative.

Before: “Weekly online assessments were conducted at Week 0, during the intervention period, and at Week 9. Weekly online assessments were conducted at Week 0, during the intervention period, and at Week 9.”

After: “Weekly online assessments were conducted at Week 0, during the intervention period, and at Week 9.”

4. The sentence “Non physician research assistants encouraged participants to use the pen consistently for their diary entries and performed minimal mental status checks during these assessments” is redundant, as the same concept is repeated afterward in the same paragraph (Intervention Group section). Therefore, it should be deleted to streamline the text.

Response: We agree with the reviewer and have removed the redundant sentence about nonphysician research assistants encouraging diary use and conducting minimal mental status checks.

Before: “Weekly online assessments were conducted at Week 0, during the intervention period, and at Week 9. Non physician research assistants encouraged participants to use the pen consistently for their diary entries and performed minimal mental status checks during these assessments.”

After: “Weekly online assessments were conducted at Week 0, during the intervention period, and at Week 9.”

5. In what manner was the viewing of the videos organized for the control group? Was a schedule in place, or were the participants free to watch the videos at their own discretion? Furthermore, how was the actual viewing of the videos ascertained?

Response: Thank you for this important question. We have revised the Methods section to clarify that while there was no formal schedule imposed for video viewing, research assistants did confirm and record whether participants had viewed the assigned video content during each weekly assessment. The sentence “Participants were free to view the videos at their own discretion, without a predefined schedule. However, research assistants confirmed and recorded whether participants had viewed the assigned video content during each assessment session.” was added in the Control Group section.

Before: “The control group received general mental health information via the Yokohama City University child psychiatry department’s website, ‘Oyako-no Kokoro-no Tomarigi’ (Appendix). This website provides educational resources about common mental health conditions in children and adolescents through easy-to-understand videos and text explanations specifically designed for young people. The video content features conversations between teddy bear and rabbit avatars discussing common mental health symptoms and concerns in children and adolescents, followed by child-friendly explanations from a child psychiatrist. Topics covered in these educational videos include: suicidal thoughts, lack of energy/motivation, anxiety, isolation and loneliness, obsessive worrying, attention difficulties, self-harm behaviors, sleep problems, and auditory hallucinations. The child psychiatrist appearing in these videos is one of the authors of this study (JF). The website also contains separate sections with mental health resources for children and families, including multiple Q&A entries about children’s mental health issues. These materials are purely informational and educational in nature, rather than providing interactive or personalized therapeutic interventions. Participants were free to view the videos at their own discretion, without a predefined schedule. However, research assistants confirmed and recorded whether participants had viewed the assigned video content during each assessment session.”

After: “The control group received general mental health information via the Yokohama City University child psychiatry department’s website, ‘Oyako-no Kokoro-no Tomarigi’ (Appendix). This website provides educational resources about common mental health conditions in children and adolescents through easy-to-understand videos and text explanations specifically designed for young people. The video content features conversations between teddy bear and rabbit avatars discussing common mental health symptoms and concerns in children and adolescents, followed by child-friendly explanations from a child psychiatrist. Topics covered in these educational videos include: suicidal thoughts, lack of energy/motivation, anxiety, isolation and loneliness, obsessive worrying, attention difficulties, self-harm behaviors, sleep problems, and auditory hallucinations. The child psychiatrist appearing in these videos is one of the authors of this study (JF). The website also contains separate sections with mental health resources for children and families, including multiple Q&A entries about children’s mental health issues. These materials are purely informational and educational in nature, rather than providing interactive or personalized therapeutic interventions. Participants were free to view the videos at their own discretion, without a predefined schedule. However, research assistants confirmed and recorded whether participants had viewed the assigned video content during each assessment session.”

6. In my personal view, the use of an active control group would have been a valuable approach, for instance, by comparing two distinct chatbots providing different types of therapy, the evaluation of which would have determined which one would prove to be more efficacious in terms of symptoms improvement. This approach would have ensured that both groups received a therapeutic intervention and could have provided additional information in terms of engagement and usability. The authors stated that the design they chose “reflects the real-world experience of many psychiatric waiting list patients in Japan,” but as they also declared, “the lack of timely intervention can exacerbate symptoms and increase the risk of severe outcomes.” Therefore, given such a risk, my question is: what is the rationale behind the authors’ decision to employ a passive control group?

Response: Thank you for this thoughtful suggestion. We agree that an active control group, such as a comparison between two therapeutic chatbots, could offer richer insights regarding engagement and efficacy.

However, our extensive prestudy evaluation revealed that emol was the only app meeting all essential criteria: (1) evidence-based therapeutic framework (ACT), (2) age-appropriate design for adolescents, (3) availability for clinical research, and (4) cost-effective access for research purposes. We conducted systematic reviews of Japanese mental health apps and interviewed multiple developers, but no comparable alternative was identified that met these combined requirements. No other chatbot meeting these criteria was identified during our review. Therefore, we adopted a passive control condition to reflect the current standard experience for patients on psychiatric waiting lists in Japan, where no structured digital intervention is provided. We have added a clarification: "While an active control group could have offered more rigorous comparison, we selected a passive control condition due to practical constraints. At the time of study planning, emol was the only adolescent-appropriate AI chatbot in Japan that integrated evidence-based psychological content (ACT), had a suitable user interface, and was available for research use. No other comparable tool was identified. Thus, we chose a passive control to reflect the real-world conditions in Japan, where patients on psychiatric waiting lists typically receive only basic informational support." in the Discussion section.

Before: "This study may have unintentionally targeted a population less receptive to alternative digital interventions. Families who had already secured an upcoming psychiatric appointment may have seen little value in participating in a study involving digital interventions, preferring instead to wait for their scheduled in-person consultation. For these families, traditional in-person care may have appeared more reassuring, especially given the severity of the patient's symptoms. Previous research on social influences in mental health service-seeking behavior among young people suggests that family is often the primary influence in choosing in-person services, whereas young people themselves tend to make decisions regarding online services [4]. Another study has also found that parents often seek informal support for their children's mental health concerns initially, only turning to professional services as issues become more severe [6]. Additionally, patients with severe symptoms or their families often prefer in-person consultations over digital interventions, perceiving in-person care as more reliable and suitable for managing serious symptoms [7]. Therefore, patients and families may value the familiarity and perceived efficacy of traditional, in-person care as a more reliable or reassuring option compared to digital alternatives. This preference likely contributed to the reluctance toward digital solutions observed in this study. Engaging patients and families earlier in the mental health care process—before they have secured traditional clinical appointments—might improve receptiveness to digital options. While an active control group could have offered more rigorous comparison, we selected a passive control condition due to practical constraints. At the time of study planning, emol was the only adolescent-appropriate AI chatbot in Japan that integrated evidence-based psychological content (ACT), had a suitable user interface, and was available for research use. No other comparable tool was identified. Thus, we chose a passive control to reflect the real-world conditions in Japan, where patients on psychiatric waiting lists typically receive only basic informational support."

After: "This study may have unintentionally targeted a population less receptive to alternative digital interventions. Families who had already secured an upcoming psychiatric appointment may have seen little value in participating in a study involving digital interventions, preferring instead to wait for their scheduled in-person consultation. For these families, traditional in-person care may have appeared more reassuring, especially given the severity of the patient's symptoms. Previous research on social influences in mental health service-seeking behavior among young people suggests that family is often the primary influence in choosing in-person services, whereas young people themselves tend to make decisions regarding online services [4]. Another study has also found that parents often seek informal support for their children's mental health concerns initially, only turning to professional services as issues become more severe [6]. Additionally, patients with severe symptoms or their families often prefer in-person consultations over digital interventions, perceiving in-person care as more reliable and suitable for managing serious symptoms [7]. Therefore, patients and families may value the familiarity and perceived efficacy of traditional, in-person care as a more reliable or reassuring option compared to digital alternatives. This preference likely contributed to the reluctance toward digital solutions observed in this study. Engaging patients and families earlier in the mental health care process—before they have secured traditional clinical appointments—might improve receptiveness to digital options. While an active control group could have offered more rigorous comparison, we selected a passive control condition due to practical constraints. At the time of study planning, emol was the only adolescent-appropriate AI chatbot in Japan that integrated evidence-based psychological content (ACT), had a suitable user interface, and was available for research use. No other comparable tool was identified. Thus, we chose a passive control to reflect the real-world conditions in Japan, where patients on psychiatric waiting lists typically receive only basic informational support."

7. The concept expressed in the sentence "Another patient refused participation due to concerns about the diary entry, and the third patient was excluded after starting therapy at another facility" is also conveyed in the preceding sentence (in the Results paragraph). It is recommended that one of the two sentences be deleted.

Response: We thank the reviewer for identifying the redundancy in our description of patient enrollment. We agree that the two sentences contained overlapping information about the same two patients. The sentence, "Another patient declined participation due to concerns about diary recording, and the third patient was excluded after beginning medication at another facility." was removed.

Before: "Among the three patients who completed the informed consent process, one participant (a female adolescent) provided consent but subsequently withdrew from the study. The

participant's family initially contacted the research team on the scheduled day of the first online session, stating: 'This morning, she became panic-stricken and is now unable to participate. Although it is the day of the appointment, would it be possible to cancel? I sincerely apologize for the inconvenience caused after all your preparations.' In a follow-up message, the family elaborated: 'She expressed anxiety about the online interview, making it impossible to proceed. We had hoped that engaging in this activity might help her develop a more positive outlook, but perhaps it was still too challenging for her.' The other two patients who completed the informed consent process either declined participation due to concerns about diary recording requirements or were excluded after beginning medication at another facility. Another patient declined participation due to concerns about diary recording, and the third patient was excluded after beginning medication at another facility. Consequently, no evaluable data were obtained in this study."

After: "Among the three patients who completed the informed consent process, one participant (a female adolescent) provided consent but subsequently withdrew from the study. The participant's family initially contacted the research team on the scheduled day of the first online session, stating: 'This morning, she became panic-stricken and is now unable to participate. Although it is the day of the appointment, would it be possible to cancel? I sincerely apologize for the inconvenience caused after all your preparations.' In a follow-up message, the family elaborated: 'She expressed anxiety about the online interview, making it impossible to proceed. We had hoped that engaging in this activity might help her develop a more positive outlook, but perhaps it was still too challenging for her.' The other two patients who completed the informed consent process either declined participation due to concerns about diary recording requirements or were excluded after beginning medication at another facility. Consequently, no evaluable data were obtained in this study."

References

1. Ennis E. Peer review of "Challenges in Implementing a Mobile AI Chatbot Intervention for Depression Among Youth on Psychiatric Waiting Lists: Randomized Controlled Study Termination Report". JMIRx Med 2025;6:e82071. [doi: 10.2196/82071]
2. Fujita J, Yano Y, Shinoda S, et al. Challenges in implementing a mobile AI chatbot intervention for depression among youth on psychiatric waiting lists: randomized controlled study termination report. JMIRx Med 2025;6:e70960. [doi: 10.2196/70960]
3. Ambrosio MDG. Peer review of "Challenges in Implementing a Mobile AI Chatbot Intervention for Depression Among Youth on Psychiatric Waiting Lists: Randomized Controlled Study Termination Report". JMIRx Med 2025;6:e82073. [doi: 10.2196/82073]
4. Rickwood DJ, Mazzer KR, Telford NR. Social influences on seeking help from mental health services, in-person and online, during adolescence and young adulthood. BMC Psychiatry 2015 Mar 7;15:40. [doi: 10.1186/s12888-015-0429-6] [Medline: 25886609]
5. Tosti B. Peer review of "Challenges in Implementing a Mobile AI Chatbot Intervention for Depression Among Youth on Psychiatric Waiting Lists: Randomized Controlled Study Termination Report". JMIRx Med 2025;6:e82074. [doi: 10.2196/82074]
6. Sawrikar V, Van Dyke C, Smith Slep AM. The Ws of parental help-seeking: when, where, and for what do parents seek help for child mental health. Child Psychiatry Hum Dev 2024 Mar 20. [doi: 10.1007/s10578-024-01683-5] [Medline: 38507021]
7. Apolinário-Hagen J, Harrer M, Kählke F, Fritsche L, Salewski C, Ebert DD. Public attitudes toward guided internet-based therapies: web-based survey study. JMIR Ment Health 2018 May 15;5(2):e10735. [doi: 10.2196/10735] [Medline: 29764797]

Abbreviations

ACT: acceptance and commitment therapy
AI: artificial intelligence

Edited by S Amal; submitted 08.08.25; this is a non–peer-reviewed article;accepted 08.08.25; published 05.09.25.

Please cite as:
Fujita J, Yano Y, Shinoda S, Sho N, Otsuki M, Suda A, Takayama M, Moroga T, Yamaguchi H, Ishii M, Miyazaki T
Authors' Response to Peer Reviews of "Challenges in Implementing a Mobile AI Chatbot Intervention for Depression Among Youth on Psychiatric Waiting Lists: Randomized Controlled Study Termination Report"
JMIRx Med 2025;6:e82083
URL: https://xmed.jmir.org/2025/1/e82083
doi:10.2196/82083

Authors' Response to Peer Reviews of "Perception and Impact of White Spot Lesions in Young People Undergoing Orthodontic Treatment and Their Guardians: Protocol for a Mixed Methods Study"

Amaar Obaid Hassan[1], BSc, BDS, MPH; Janine Doughty[2], BDS, MDPH, DDPH RCS Eng, PG Cert Clin Res, PGCAP, PhD; Jayne Harrison[3], BDS, MDentSci, PhD

[1]Department of Orthodontics, School of Dentistry, Liverpool University, Pembroke Place, Liverpool, United Kingdom

[2]School of Dentistry, University of Liverpool, Liverpool, United Kingdom

[3]Orthodontic Department, Liverpool University Dental Hospital, Liverpool, United Kingdom

Corresponding Author:
Amaar Obaid Hassan, BSc, BDS, MPH
Department of Orthodontics, School of Dentistry, Liverpool University, Pembroke Place, Liverpool, United Kingdom

Related Articles:

Companion article: http://preprints.jmir.org/preprint/60213

Companion article: https://med.jmirx.org/2025/1/e80143

Companion article: https://med.jmirx.org/2025/1/e80140

Abstract

(JMIRx Med 2025;6:e80139) doi:10.2196/80139

KEYWORDS

orthodontics; white spot lesions; fixed appliances; dentistry

This is the authors' response to peer-review reports for "Perception and Impact of White Spot Lesions in Young People Undergoing Orthodontic Treatment and Their Guardians: Protocol for a Mixed Methods Study."

Round 1 Review

Reviewer A [1]

General Comments

This paper [2] appears to me to be well-written and adequately cited. I believe that this paper will contribute to the literature once the study commences and the data are collected and analyzed. However, I do have some questions/concerns regarding the study design and potential data analysis that I have included in my comments below. I would like the authors of this paper to review these comments/recommendations and to either implement them as they see fit or justify why they believe they do not need to.

Specific Comments

Major Comments

B1. Line 170, you state that this study is purely descriptive, so a power analysis is not required. How will you control for confounding variables such as cultural beliefs which may be over- or underrepresented in your participant pool? Additionally, how will you ensure that your participant demographics allow for the generalization of this paper's findings to patient populations outside of Liverpool?

Response: Following the advice of the peer review, we have undertaken a pilot study including 20 people, and this has enabled us to complete a power calculation and sample size, meaning we are able to provide statistically significant data of the representative sample. However, if we want a representative sample from patients receiving routine orthodontic treatment in our department, with 11 providers and about 50 patients each, this gives a population of 550 and with a 95% confidence level and error margin of 5%, we would need 226 respondents.

B2. Line 203, you mention that sampling will be based on age, gender, ethnicity, etc. However, Table 2 does not mention ethnicity. Could you edit Table 2 to mention ethnicity or edit

Line 203 to remove ethnicity. I would recommend editing the table because I believe the participant demographics to be important, especially since different cultures may approach esthetics and health beliefs differently. This concern regarding culture connects with major comment 1.

Response: The table has been amended to add ethnicity, meaning that at least 4 of 12 participants will be required to be from a minority ethnic background, thank you.

Minor Comments

B3. Line 129, "Sponsorship will be sort from...," please change to "Sponsorship will be sought from..."

Response: As now approved, we have changed this to: "Sponsorship has been obtained from Liverpool University Hospitals NHS Foundation Trust (UoL001871)."

B4. Line 167, you state that a sample size of 200 respondents is sufficient for Part 1 of the study. Could you justify this estimate in a more thorough way other than stating that it is a "pragmatic estimation?"

Response: We have undertaken a sample size calculation following a pilot study where we invited 20 participants to provide feedback on the questionnaire. Using this data, we were able to complete an analysis of the statistical difference and thus justify the sample size using a power calculation. See above response to B1.

B5. Line 173, you state that participants will be contacted on the same day as their orthodontic appointment. Will this be before or after the appointment? Will participants be compensated for their time? How will you ensure that participants' rights are respected and that they do not feel pressured into participating?

Response: Participants will be compensated for their time, and the statement "They will have free choice over whether they wish to take part and will be able to read the participant information leaflet during their appointment or after. Should they wish to take part, then they will be reimbursed for their time, with a £10 electronic voucher for the questionnaire, and £25 for the qualitative research" has been added. In the pilot, 25% of people declined to take part even with an offer of an electronic voucher as reimbursement. We have amended the timeline following the pilot study and have increased the time for recruitment to 12 months, meaning there will be ample time for recruitment and no pressure for the research team. Participants will be contacted while they attend their orthodontic appointment.

B6. Line 427, "or childs name?" please change to "or child's name?"

Response: We are unable to identify the error you have stated.

Reviewer AH [3]

C1. The abstract of this paper must be revised. "Several studies explore the prevention and/or treatment of WSL" is inappropriate for the abstract.

Response: We have amended this to "Although there have been studies that have investigated the prevention and treatment for WSL, there remain uncertainties about what young people and their parents or guardians know or feel about them."

C2. The Methods section describes the mixed methods approach well, but the recruitment process could be elaborated. For example, how will convenience sampling be conducted to avoid bias? Including qualitative and quantitative data is well-justified, but there is no mention of how the two datasets will be integrated into the analysis. More details on how the qualitative data will expand upon the quantitative findings would strengthen the methodology.

Response: We have agreed to review the data collection after 25% of data collection has been completed. We will check age, gender, ethnicity, stage of treatment, Index of Orthodontic Treatment Need, and condition of first molar teeth and determine whether we feel this is consistent with the representative sample of the population. If it is not, we will target more underrepresented groups. One hundred is a large sample size, and many of our patients have not commenced with orthodontic treatment or are above the age of 15 and do not meet our inclusion criteria, meaning that those who do meet the inclusion criteria are likely to be contacted to take part. We have added the following statement: "The authors will review the data after 25% completion of the quantitative study to check if participants who have been recruited fit the demographics of the clinic. If any people are underrepresented at this point, then this will be identified, and more efforts will be made to recruit people from the underrepresented groups." We know from a previous audit the people who have brace treatment at the clinic are representative of the sample population for deprivation status. We have included details on how the data will be integrated, stating that "The mixed methods study design is to provide enriched data by augmenting the quantitative findings with qualitative interviews. An explanatory sequential mixed methods approach will be used, whereby the qualitative data will expand on the understanding gained from the questionnaire [4]. The diagram below (Figure 2) illustrates the different parts of the study and at what point the mixing of the data will occur. The quantitative and qualitative parts of the research will be analyzed for convergent and divergent data interpretation of the mixed methods research that compares both datasets. Figure 2. Flowchart of mixed methods design." We have also included a flowchart (see Figure 2), illustrating how the data are to be integrated. For clarity, we have also added the following statements: "Following completion of the quantitative research, we will organize meetings with the research team and patient and public involvement group to review the data and develop the interview schedule based on the findings of the first part of the research" and "Following completion of the data collection/analysis of both datasets, the results will be merged. The quantitative and qualitative data will then be compared for convergence and divergence."

C3. The sample size rationale is explained well, though stating why a power analysis is unnecessary for a descriptive study could prevent confusion.

Response: Since we have received feedback for the study, we have undertaken a power analysis and sample size calculation based on a pilot study we completed. See the response to B1.

C4. In the Results section, it would be useful to clarify how data from questionnaires and interviews will be compared and whether there is an expectation of divergence between parent and child responses.

Response: We are currently undertaking a statistical analysis on questions in the questionnaire between parents and children so that we are able to compare answers between children and parents. Questions we are looking at comparing include "before braces and after braces are removed but with WSL" photos and "how likely you think you will get WSL." We are unsure if there will be an expectation of divergence between parent and child responses; although in the pilot, we gathered information from 10 children and 10 parents/guardians, and these answers did differ (parents expected that their child was more likely to get white spot lesions [WSLs] and were more unhappy about getting WSLs compared to the children). The overall κ statistic for these questions was 0.284 (95% CI 0.029-0.539), with individual questions ranging from −0.152 to 0.98, suggesting that there was fair agreement but that there was considerable variation. This will be explored further and discussed once the study has been completed. We are expecting that the qualitative research will have similar findings to the quantitative part. We have added the following statement in the Results section: "We will also use the κ statistic to determine whether there are differences between parents'/guardians' and young people's answers to the questions in the questionnaire."

C5. The Limitations section acknowledges some important aspects, such as recruiting from only one hospital, but it does not address potential biases in self-reported data. There is also no mention of how the study will address participants' potential reluctance to report negative experiences due to social desirability bias. Expanding on these limitations and how the study will mitigate them would improve transparency.

Response: We think that we will address potential biases in self-reported data by using statistical analysis in the quantitative research to help confirm study findings, and we will also use a coding framework (NVivo) to generate codes/themes (rather than developing codes ourselves) to try to limit self-reported bias. We have also included a researcher in the team who is not a clinician to assist with recruitment and data collection and analysis. We have commented in the study that multiple people are interpreting codes/data. We have used a patient and public involvement group to develop the research and ensure that it is patient-focused, and we will continue to do this to develop a questionnaire and interview schedule to avoid leading questions. We have written up the study protocol, which has undergone peer review as part of the grant application and ethical approval processes. This will help to ensure transparency and has involved a diverse group of opinions to identify potential sources of bias. We have also published our questionnaire and interview schedule, which has been added as an appendix. With qualitative research, there is always a limitation of self-reported bias; however, we have attempted to limit this. To clarify this in the Limitations section, we have identified the following statements: "Although one of the limitations of survey and qualitative research includes the potential risk of self-reporting bias, the authors have attempted to address this by publishing the study protocol and using patient and public involvement to develop the research and help to analyze/interpret the data. The authors will publish the protocol, the questionnaire/interview schedule, and data so that readers are able to make an informed decision about the potential sources of bias. Data analysis will be reviewed by a researcher who is a nonclinician and NVivo will be used to limit self-reporting and ensure a systematic framework to coding" and "Participants also have the opportunity to review study findings to ensure that they agree with the results" and "During the qualitative research analysis, data coding and themes of transcripts will be undertaken by AOH using NVivo 12. The transcripts and codes/themes generated will be sent to a second or third researcher to confirm reliability (JH, JD, or AR)." We have attempted to address social desirability bias by asking the young participants not to discuss answers with parents/guardians as this may influence their answers. Participants are advised that the reason for separate questionnaires is so that they can answer their questions honestly and that the authors are able to compare answers. The participants can complete the questionnaire in a private room without a researcher being present. The study has used patient and public involvement throughout to ensure questions are relevant to the participants and not misleading. The following statements have been added to the Limitations section: "The authors have also attempted to address self-reporting bias by publishing the study protocol, the questionnaire/interview schedule, and data so that readers are able to make an informed decision about the potential sources of bias" and "Social desirability bias has been limited by asking participants not to discuss answers with parents/guardians as this may influence their answers. The participants are able complete the questionnaire in a private room without a researcher being present. The study will not recruit any participants who are under the clinical care of the research team involved in recruiting. Patient and public involvement will be used throughout all stages of the research to ensure questions are relevant to the participants and not misleading."

C6. The potential psychological impact of WSLs could be expanded upon in the Discussion, especially regarding how WSLs may affect patient compliance and satisfaction posttreatment.

Response: Negative association following WSLs has already been discussed in the Discussion section. We have added the following statement regarding improving compliance for preventing WSLs in the Discussion section: "Even with effective oral hygiene instruction, around half of young people do not follow the clinician's advice to improve their oral hygiene [5]. The COM-B model is presented as a tool to diagnose which of capability, opportunity, or motivation need to change for a new behaviour to take place [6]. Although interventions designed to improve oral hygiene during orthodontic treatment (including using smartphones, a toothbrushing app, visual aids, motivational interviewing, oral health reinforcements) have been looked at, it is only the use of mobile phones that have limited evidence for improving oral health during orthodontic treatment [7]. To our knowledge, trials/studies have not been undertaken to explore barriers to oral hygiene or behavioural interventions to reduce WSL formation during orthodontic treatment in young people."

C7. To expand the Discussion, the following article must be cited: Jamloo H, Majidi K, Noroozian N, et al. Effect of fluoride on preventing orthodontics treatments-induced white spot lesions: an umbrella meta-analysis. Clin Investig Orthod. April 19, 2024;83(2):53 - 60. [doi: 10.1080/27705781.2024.2342732]

Response: We have added the reference in the Introduction section.

References

1. Shaw J. Peer review of "Perception and Impact of White Spot Lesions in Young People Undergoing Orthodontic Treatment and Their Guardians: Protocol for a Mixed Methods Study". JMIRx Med 2025;6:e80143. [doi: 10.2196/80143]
2. Hassan AO, Doughty J, Harrison J. Perception and impact of white spot lesions in young people undergoing orthodontic treatment and their guardians: protocol for a mixed methods study. JMIRx Med 2025;6:e60213. [doi: 10.2196/60213]
3. Jamilian A. Peer review of "Perception and Impact of White Spot Lesions in Young People Undergoing Orthodontic Treatment and Their Guardians: Protocol for a Mixed Methods Study". JMIRx Med 2025;6:e80140. [doi: 10.2196/80140]
4. Schoonenboom J, Johnson RB. How to construct a mixed methods research design. Kolner Z Soz Sozpsychol 2017;69(Suppl 2):107-131. [doi: 10.1007/s11577-017-0454-1] [Medline: 28989188]
5. Mei L, Chieng J, Wong C, Benic G, Farella M. Factors affecting dental biofilm in patients wearing fixed orthodontic appliances. Prog Orthod 2017 Dec;18(1):4. [doi: 10.1186/s40510-016-0158-5] [Medline: 28133715]
6. Michie S, van Stralen MM, West R. The behaviour change wheel: a new method for characterising and designing behaviour change interventions. Implement Sci 2011 Apr 23;6(42):42. [doi: 10.1186/1748-5908-6-42] [Medline: 21513547]
7. Farhadifard H, Soheilifar S, Farhadian M, Kokabi H, Bakhshaei A. Orthodontic patients' oral hygiene compliance by utilizing a smartphone application (Brush DJ): a randomized clinical trial. BDJ Open 2020 Nov 20;6(1):24. [doi: 10.1038/s41405-020-00050-5] [Medline: 33298841]

Abbreviations

WSL: white spot lesion

Edited by E Meinert, T Leung; submitted 04.07.25; this is a non–peer-reviewed article;accepted 04.07.25; published 12.09.25.

Please cite as:
Hassan AO, Doughty J, Harrison J
Authors' Response to Peer Reviews of "Perception and Impact of White Spot Lesions in Young People Undergoing Orthodontic Treatment and Their Guardians: Protocol for a Mixed Methods Study"
JMIRx Med 2025;6:e80139
URL: https://xmed.jmir.org/2025/1/e80139
doi:10.2196/80139

Author's Response to Peer Reviews of "COVID-19 Pneumonia Diagnosis Using Medical Images: Deep Learning-Based Transfer Learning Approach"

Anjali Dharmik, MSc

Royal Holloway University of London, Egham Hill, Egham, United Kingdom

Corresponding Author:
Anjali Dharmik, MSc
Royal Holloway University of London, Egham Hill, Egham, United Kingdom

Related Articles:

Companion article: https://arxiv.org/abs/2503.12642v2

Companion article: https://med.jmirx.org/2025/1/e83231

Companion article: https://med.jmirx.org/2025/1/e83234

Companion article: https://med.jmirx.org/2025/1/e83236

Companion article: https://med.jmirx.org/2025/1/e75015

(JMIRx Med 2025;6:e83230) doi:10.2196/83230

KEYWORDS

computer vision; COVID-19 pneumonia diagnosis; deep learning; transfer learning; medical imaging analysis

This is the authors' response to peer-review reports for "COVID-19 Pneumonia Diagnosis Using Medical Images: Deep Learning-Based Transfer Learning Approach."

Round 1 Review

Reviewer S [1]

General Comments

This paper [2] focused on the use of artificial intelligence (AI), in particular convolutional neural networks (CNNs) for detection of COVID-19 infections in radiological imaging. The study uses a substantial dataset of over 6000 images, which enhances the reliability of the results and supports robust model training and evaluation. Leveraging well-known CNNs such as VGG16, VGG19, and ResNet-50 demonstrates a practical application of transfer learning, a widely accepted technique in deep learning for medical imaging tasks.

However, in the Background and Introduction sections, the authors focused on the importance of rapid and early diagnosis of COVID-19, thus the demand for AI CNNs for diagnosis ("traditional diagnostic methods, such as serologic tests, have limitations, including low sensitivity and longer processing times"), yet this could be achieved easily nowadays with lateral flow devices or rapid antigen tests. Using computed tomography or X-rays to screen COVID-19 is far too expensive and time consuming compared to lateral flow devices or rapid antigen tests.

I believe the author was referring to the use of AI CNNs to differentiate COVID-19 pneumonia from other causes of pneumonia. Diagnosis of COVID-19 infection (which is usually mild and self-limiting) is totally different from COVID-19 pneumonia (which might require hospitalization and medical interventions). The authors might consider changing the title of the manuscript to "COVID 19 Pneumonia Diagnosis Analysis Using Transfer Learning–Deep Learning." Similarly, for section 3.1, "COVID-19 Pneumonia Diagnosis Using Deep Learning" would be more appropriate than "COVID-19 Diagnosis Using Deep Learning."

In addition, the Related Work section is brief and lacks depth. It does not sufficiently review existing medical studies on deep learning for COVID-19 pneumonia diagnosis, making it less comprehensive.

Response: I updated the title to "COVID-19 Pneumonia Diagnosis Using Medical Images: Deep Learning-Based Transfer Learning Approach" and added more details on research into COVID-19 pneumonia diagnosis in the current situation and existing studies.

Specific Comments

Major Comments

1. *Since this is a medical journal, medical terms are encouraged, for example,* anosmia *to replace* loss of smell*;* ageusia *to replace* loss of taste.

Response: I added more detail about COVID-19 pneumonia diagnosis and the current situation.

2. Quite a significant number of references were not medical related, but related to AI or computer science. I would suggest the authors visit PubMed to search for more medical-related references. I cannot suggest any particular references to avoid conflicts of interest with certain groups of authors and to avoid self-citation.

Response: I added and updated the references to include medical studies and technical studies.

3. The author detailed the AI CNN mechanism, yet the features of computed tomography or X-rays that were focused on were not mentioned. Was ground glass appearance the main target, or was it other features like cavitation, extent of lung involvement, or superior location? It would be more valid to evaluate various features targeted by the AI, instead of mentioning how it works.

Response: Thank you for the suggestion. While the study used Grad-CAM to visualize model attention, highlighting features like ground glass opacities and bilateral lower-lobe involvement, we acknowledge that explicitly evaluating a wider range of radiological features (eg, cavitation, extent, and location) would strengthen the clinical validity, and these will be considered in future work.

Minor Comments

4. The author cited many different online references, yet the links or URLs were not available for readers to refer to. I would suggest adding the cited reference sources back for reviewers to assess the appropriateness of the citation, such as references 26 to 28, and for the benefit of readers. For example, reference 9 is not searchable on the internet.

Response: I updated the references.

5. In section 1.1, "At that point, there have been 98 confirmed cases and no reported deaths in 18 countries outside China..." Please add a reference citation for this factual statement.

Response: I updated this.

6. In section 1.1, "As of 28 April 2020, 63% of worldwide mortality from the virus was from the Region..." Please clarify the "Region."

Response: I updated this.

7. In section 1.3, "Motivation to try to COVID-19 Diagnosis," the English could be further polished, for example, "Motivation-to-try to COVID-19 Diagnosis" or "Motivation to try towards COVID-19 Diagnosis."

Response: I updated this.

8. Computed tomography*, instead of* computer tomography*, is the proper term in section 3.1.*

Response: I updated this.

9. Abbreviations need not be spelled out again after their first use in the main text. For example, "computer tomography (CT)" in section 3.1 can just be "CT," since CT has been defined already.

Response: I updated this.

*10. Please be consistent with reference citation formatting; various formats are used in the reference list, for example, "[20] Z. Wu et al Unsupervised feature learning via non-parametric instance discrimination. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 3733 - 3742, 2018"; "[24] Md. Islam, F. Karray, R. Alhajj, and J. Zeng. A review on deep learning techniques for the diagnosis of novel coronavirus (covid-19). IEEE Access, vol. 9, pp. 30551 - 30572, 2021. doi: 10.1109/ACCESS.2021.3058537"; and "[29] Mohammed K. Hassan, Ali I. El Desouky, Sally M. Elghamrawy, and Amany M. Sarhan. A hybrid real-time remote monitoring framework with nb-woa algorithm for patients with chronic diseases. http://doi.org/10*https://doi.org/10.*1016/j.future.2018.10.021, 2019. Future Generation Computer Systems, Volume 93, Pages 77 - 95, ISSN 0167 - 739X."*

Response: I updated this.

11. To further improve the manuscript, please consider adding figures or tables showing the appearance of COVID-19 versus normal samples. Add to the Limitations section a discussion of potential biases (eg, dataset origin) or generalizability issues (eg, applicability to new variants) to demonstrate critical reflection

Response: I updated this.

Reviewer AA [3]

General Comments

This manuscript [2] describes a transfer-learning approach using pretrained convolutional neural networks (VGG16, VGG19, ResNet-50) for binary COVID-19 detection on chest X-ray and computed tomography images. Overall, it tackles a timely problem and reports high accuracy (>97%), but several methodological and reporting issues limit confidence in the findings and their reproducibility.

Specific Comments

Major Comments

1. Lack of clinical validation: no in vivo or clinical ground-truth data are provided. The model's >97% accuracy is based solely on public datasets; it's unclear how it performs on real-world, heterogeneous clinical images.

Response: I updated this.

2. Overfitting and hyperparameter tuning: identical performance across 5 hyperparameter settings for VGG16 suggests under- or overfitting. No learning curves or regularization impact analyses are shown to substantiate robustness claims.

Response: I updated this.

3. Model comparison baseline: no comparison against simple baselines (eg, logistic regression on hand-crafted features) or recent literature benchmarks is provided, making it difficult to evaluate novelty and real gain.

Response: Thank you for the observation. In our study, we used a baseline AlexNet (a convolutional neural network–inspired architecture) to benchmark performance against more advanced transfer learning models. Our primary focus was on evaluating the effectiveness of various transfer learning approaches.

Minor Comments

4. Repeated headings: "Integration into Mobile/Cloud-based Platform" appears twice in section 1; please consolidate.

Response: I updated this.

5. Typographical and formatting errors: multiple sentences start without capitalization (eg, "we reviewing to the difference…") and several references lack publication details (eg, [27,28] list only URLs).

Response: I updated this.

Reviewer AB [4]

General Comments

This manuscript [2] investigates the application of deep learning, particularly transfer learning using VGG16, VGG19, and ResNet-50, for diagnosing COVID-19 through computed tomography and X-ray images. The topic is important and timely, especially considering the enduring threat of COVID-19 variants and the burden on global health care systems. The author demonstrates technical familiarity with deep learning techniques, model tuning, and performance evaluation. However, there are areas where the study could be improved to enhance its rigor, clarity, and impact.

Specific Comments

Major Comments

1. Dataset description and bias: the paper mentions using a dataset of 6259 images (4651 COVID-19 cases and 1608 normal cases). However, there is no discussion on potential biases in the dataset, such as the source of the images, demographic diversity (age, gender, and geographic location), or the balance between COVID-19 and normal cases. Addressing these aspects would strengthen the validity of the results. I suggest that the author include a detailed description of the dataset, including sources, demographic information, and steps taken to mitigate bias, and consider discussing the imbalance in the dataset and how it might affect model performance.

Response: I updated this.

2. Comparative analysis with existing methods: while the paper reports high accuracy (97.73%) for the proposed models, it lacks a comprehensive comparison with other state-of-the-art methods or baseline models. This makes it difficult to assess the novelty and superiority of the proposed approach. I suggest that the author add a comparative table or section that contrasts the performance of VGG16, VGG19, and ResNet-50 with other recent studies or baseline models and highlight the unique contributions of this work.

Response: I updated this.

3. Clinical relevance and practical deployment: the study focuses on technical performance metrics but does not discuss the clinical applicability of the models. For instance, how would these models integrate into real-world health care settings? What are the potential challenges (eg, computational resources, interpretability for clinicians)? I suggest that the author expand the discussion on clinical relevance, including limitations and practical considerations for deployment in health care systems.

Response: I updated the paper to mention potential challenges, discuss clinical relevance, mention limitations, and discuss practical considerations for deployment in health care systems.

4. Language and grammar: the manuscript needs extensive language editing. There are frequent grammatical issues, awkward phrasing (eg, "the 1608 belong to healthy people"), and repetition. A professional edit is highly recommended to improve readability and flow.

Response: I updated this.

5. Figures and tables: several figures (eg, confusion matrices, loss/accuracy curves) are referenced but lack sufficient clarity, labeling, or captions. Figures 4 to 8 must be embedded clearly within the results discussion and interpreted to guide the reader. Ensure figures are high resolution and correctly formatted.

Response: I updated this.

6. Overstatement of results: the paper claims high performance (97.73% accuracy), yet offers little discussion on external validity or overfitting risks. Since cross-validation was performed on a relatively small dataset, these results may not generalize well. The author should tone down claims and discuss limitations.

Response: I added a detailed discussion of overfitting risks, cross-validation, datasets, and results.

7. Dataset description and ethics: while the dataset is described as publicly available, the manuscript lacks ethical approval or justification. Clarify whether ethical clearance was required. Also, organize the dataset description into a single, detailed section including data sources, balance between classes, preprocessing applied, and augmentation steps.

Response: I updated the paper to describe collection of the data sources and mention the processing steps.

8.Evaluation metrics and statistical rigor: the paper heavily relies on accuracy, sensitivity, specificity, and F1-score, but fails to report CIs or conduct statistical tests to validate performance differences between models. Including receiver operating characteristic area under the curve values and visualizations would also strengthen the evaluation.

Response: I included receiver operating characteristic area under the curve values and added a visualization to the Results section.

9. Novelty and contribution not clearly established: while the paper uses popular convolutional neural network architectures, there is no clear indication of what is novel in this study compared to the extensive body of existing work using these same models on similar datasets. What distinguishes this work? Is it the dataset size, preprocessing technique, tuning strategy, or model ensemble?

Response: I updated these details.

Minor Comments

10. Hyperparameter tuning details: the paper describes hyperparameter tuning but does not explain the rationale behind the selected ranges (eg, learning rate and batch size). A brief justification for these choices would improve reproducibility. I suggest adding a sentence or two explaining why the specified ranges for hyperparameters were chosen.

Response: I added a discussion of the hyperparameter tuning.

11. Use consistent terminology throughout (eg, "deep learning model" versus "CNN-based model").

Response: I updated this.

12. Data augmentation techniques: these are described generically. Specify which augmentations were applied and how frequently. Were augmentation parameters validated?

Response: I updated this discussion with more details.

13. Please structure the abstract under clear headings, Background, Objective, Methods, Results, and Conclusion, to aid clear reading and comprehension.

Response: I updated this.

Round 2 Review

Reviewer S [1]

Specific Comments

Major Comments

Some parts of the manuscript[1] used extensive bulleted lists; paragraphs should be used in the manuscript's main text. If the author deems bullet points more appropriate for the content, the author could format lists as tables.

Response: I rewrote the bullet points as full paragraphs.

References

1. Au SCL. Peer review of "COVID-19 Pneumonia Diagnosis Using Medical Images: Deep Learning-Based Transfer Learning Approach". JMIRx Med 2025;6:e83231. [doi: 10.2196/83231]
2. Dharmik A. COVID-19 pneumonia diagnosis using medical images: deep learning-based transfer learning approach. JMIRx Med 2025;6:e75015. [doi: 10.2196/75015]
3. Odezuligbo I. Peer review of "COVID-19 Pneumonia Diagnosis Using Medical Images: Deep Learning-Based Transfer Learning Approach". JMIRx Med 2025;6:e83234. [doi: 10.2196/83234]
4. Ndezure E. Peer review of "COVID-19 Pneumonia Diagnosis Using Medical Images: Deep Learning-Based Transfer Learning Approach". JMIRx Med 2025;6:e83236. [doi: 10.2196/83236]

Edited by F Wu; submitted 29.08.25; this is a non–peer-reviewed article;accepted 29.08.25; published 26.09.25.

Please cite as:
Dharmik A
Author's Response to Peer Reviews of "COVID-19 Pneumonia Diagnosis Using Medical Images: Deep Learning-Based Transfer Learning Approach"
JMIRx Med 2025;6:e83230
URL: https://xmed.jmir.org/2025/1/e83230
doi:10.2196/83230

Author's Response to Peer Reviews of "Real-Time Health Monitoring Using 5G Networks: Deep Learning–Based Architecture for Remote Patient Care"

Iqra Batool, MSEng

Department of Computer Science, Western University, 1151 Richmond St, London, ON, Canada

Corresponding Author:
Iqra Batool, MSEng
Department of Computer Science, Western University, 1151 Richmond St, London, ON, Canada

Related Articles:

Companion article: https://arxiv.org/abs/2501.01027v1

Companion article: https://med.jmirx.org/2025/1/e83423

Companion article: https://med.jmirx.org/2025/1/e83424

Companion article: https://med.jmirx.org/2025/1/e70906

(JMIRx Med 2025;6:e83473) doi:10.2196/83473

KEYWORDS

5G; real-time patient monitoring; vital signs; prediction; deep learning; machine learning

This is the authors' response to peer-review reports for "Real-Time Health Monitoring Using 5G Networks: Deep Learning–Based Architecture for Remote Patient Care."

Round 1 Review

Reviewer S [1]

We thank Reviewer S for their constructive feedback and positive evaluation of our work [2]. Below, we provide a detailed response to each comment and indicate where these concerns have been addressed in our revised manuscript.

The combination of a convolutional neural network/long short-term memory model with 5G ultra-reliable low latency communication enables real-time monitoring with high accuracy and low latency. Achieving 96.5% accuracy for vital sign prediction demonstrates the effectiveness of the proposed model.

Response: We appreciate the reviewer's recognition of our system's performance. The 96.5% accuracy is detailed in section 5 (Results and Analysis), with comprehensive performance metrics shown in Table 1, demonstrating mean absolute error values of 1.82%, 2.14%, and 1.95% for heart rate, blood pressure, and respiratory rate, respectively.

While tested on 1000 patients, analysis of its scalability to larger populations with diverse demographics would improve generalizability.

Response: We have extensively addressed scalability concerns in section 5.5 (System Scalability and Performance Analysis). Our analysis includes the following:

- Computational scalability; linear scaling up to 2000 concurrent patients with graceful degradation beyond this threshold
- Network scalability; support for up to 1000 high-priority patients (intensive care/critical care) and 4000 standard-priority patients simultaneously
- Mathematical modeling; scalability relationship modeled using equation 26
- Diverse patient populations; performance validated across critical care (96.5%), postoperative (95.8%), and general ward patients (97.2%)

The use of attention mechanisms in the long short-term memory component improves the system's ability to model dependencies in continuous vital sign monitoring.

Response: The attention mechanism implementation is detailed in section 3.2 (Deep Learning Framework) with mathematical formulations in equations 10 and 11. The 4-head attention mechanism was optimized through Bayesian optimization, as described in section 4.3 (Model Development).

A more detailed comparison with state-of-the-art remote monitoring systems, including their architectures and limitations, would strengthen the claims.

Response: We have provided comprehensive comparisons in multiple sections:

- Section 4.2—detailed baseline comparison with 3 established systems (ConventionalCare RPM Platform, EdgeMed Smart Monitoring, and NextGen 5G Health Platform)
- Section 5.7—comparative analysis with performance metrics in Table 4
- Statistical validation—statistical significance testing in Table 5, with paired *t*-tests confirming improvements ($P<.001$)
- Architecture details—each baseline system includes mathematical formulations and operational specifications

Since patient data is transmitted over 5G networks, an evaluation of encryption techniques, data integrity measures, and compliance with health care regulations (eg, the Health Insurance Portability and Accountability Act and the General Data Protection Regulation) should be included.

Response: Security and privacy are comprehensively addressed in section 4.4 (Security Architecture and Data Protection):

- Encryption—AES-256 encryption with mathematical formulation (equation 24)
- Privacy-preserving techniques–differential privacy implementation (equation 25)
- Regulatory compliance—Health Insurance Portability and Accountability Act and General Data Protection Regulation compliance with specific implementation details
- Network security—5G network slicing with isolated communication channels
- Data protection—end-to-end encryption, public key infrastructure management, and role-based access control

Investigating performance under network congestion, packet loss, or fluctuations in 5G coverage would ensure system reliability.

Response: Network robustness is thoroughly evaluated in section 5.4 (Network Robustness and Reliability Assessment):

- Network congestion—performance maintained at 96.1% accuracy under 50% capacity, 95.3% at 75% congestion
- Packet loss tolerance—system maintains 96.2% accuracy with 1% packet loss, 94.8% with 5% loss
- Coverage fluctuation—automatic fallback mechanisms to 4G with monitoring continuity maintained
- Reliability modeling—mathematical formulation in equation 21

Reviewer BM [3]

We thank reviewer BM for their careful review and detailed feedback. All identified issues have been addressed in our revised manuscript. Below is our point-by-point response indicating where each concern has been resolved.

Major Comments

Table 3 is not referenced nor commented on in the text. You should add a paragraph explaining the table or delete it.

Response: Table 3 is now properly referenced and explained in section 5.3 (System Latency Analysis). A detailed paragraph has been added explaining the latency breakdown across different processing stages, highlighting that the total pipeline latency of 14.4 ms meets real-time clinical monitoring requirements. Location: section 5.3, paragraph discussing end-to-end latency measurements.

Table 5 compares the system performance with 3 other systems, A, B, and C, but those systems are never described. They must be commented on in order to compare results.

Response: The 3 baseline systems are comprehensively described in section 4.2 (Baseline Comparison Systems). Additionally, clear cross-references have been added in section 5.7 (Comparative Analysis) directing readers to these detailed descriptions before presenting the comparison results. Locations: detailed descriptions in section 4.2; cross-references added in section 5.7, before Table 4.

Minor Comments

Equation 1 has no label (1) and it is defined twice.

Response: All equations have been properly numbered sequentially throughout the manuscript. Equation 1 now appears only once, with proper labeling, and all subsequent equations follow the correct numerical sequence. Location: throughout the manuscript, starting with equation 1 in section 3.2.

Figure 4 should be placed after it is called out.

Response: Figure 4 has been repositioned to appear immediately after its first call-out, following standard manuscript formatting guidelines. Location: section 5.1, after the first mention of the performance timeline.

On page 6, there is a sentence in square brackets.

Response: All editorial comments and square bracket notations have been removed from the manuscript. The document has been thoroughly reviewed to ensure no editorial marks remain. Location: page 6 content has been cleaned and integrated into the proper text.

Correct the sentence "Figure 4illustrates..." The number 4 and the word "illustrates" are too close.

Response: The spacing error has been corrected to read "Figure 4 illustrates..." with proper spacing between the figure number and text. All similar formatting issues throughout the document have been resolved. Location: section 5.1, Performance Evaluation subsection.

Table 5 is called out before Table 4. Consequently, they should be switched.

Response: Tables have been reordered to match their sequence in the text. All table numbers and corresponding mentions have been updated accordingly to maintain proper numerical order. Location: tables now appear in correct sequence in section 5.

The sentence "Table V System Comparison..." seems to be a figure description instead of part of the text. It makes no sense in the place it is located.

Response: All table captions have been properly formatted and positioned according to journal guidelines. Table descriptions have been moved from body text to appropriate caption format. Location: all tables in section 5 now have properly formatted captions.

The text "(P ! .001)" I presume should be "(P<.001)"

Response: All instances of incorrect mathematical notation such as "p ! 0.001" have been corrected to "*P*<.001". The entire manuscript has been reviewed for mathematical symbol accuracy. Location: section 5.7, Table 5, and associated statistical analysis text.

Round 2 Review

Reviewer BM [3]

First of all, we would like to thank the reviewers for their valuable reviews. We have addressed all the comments, and our detailed responses are below.

There are some equations with no defined parameters. In equation 16, what are Pij and xij? In equation 17, what is N? In equation 18, what are Bi, Cj, and M? In equation 19, what is Lu? They must be defined.

Response: We acknowledge this critical oversight and have thoroughly revised all equations to include complete parameter definitions. The corrections are as follows, under the Resource Allocation heading:

"The resource allocation for the healthcare slice is optimized using:

$min\sum_i\sum_j P_{ij}x_{ij}$ (16)

subject to:

$\sum_j x_{ij}=1, \forall i \in N$ (17)

Where:

P_{ij}=power consumption (Watts) when patient *i* is assigned to server *j*

x_{ij}=binary resource allocation variable (1 if patient *i* is assigned to server *j*, *0* otherwise)

N=set of all patients requiring monitoring, N={1,2,...,n}

M=set of available edge computing servers, M={1,2,...,m}

B_i=bandwidth requirement of patient *i* (Mbps)

C_j=computational capacity of server *j* (operations per second)

The resource allocation optimization considers four critical system parameters. Power consumption P_{ij} affects overall energy efficiency and operational costs of the monitoring infrastructure. The binary allocation variable x_{ij} governs the distribution of computational resources across the network, ensuring each patient is assigned to exactly one processing server. Bandwidth requirements B_i determine the communication overhead for transmitting vital sign data from each patient, while capacity constraints C_j ensure the system operates within the feasible computational limits of each edge server.

Constraint (17) ensures that each patient is assigned to exactly one server, preventing resource conflicts and ensuring complete coverage. Constraint (18) guarantees that the total computational load assigned to any server does not exceed its processing capacity, maintaining system stability and response time requirements."

How are weights wu, wr, and wl calculated or estimated? What are their chosen values? The final performance could change depending on the selection of these parameters, as you are giving more importance to one parameter or another.

Response: This is an important methodological question that we have addressed by adding a dedicated subsection, "Weight Parameter Determination," under the Resource Allocation section:

"$P(i)=w_uU_i+w_rR_i+w_lL_i$ (20)

where:

U_i is the urgency factor

R_i is the reliability requirement

L_i is the latency requirement

w_u,w_r,w_l are corresponding weights

Real-time latency monitoring and dynamic route optimization further enhance the system's reliability and performance through continuous assessment shown in equation (21):

$R(t)=(1–P_e)(1–P_l)(1–P_u)$ (21)

where:

P_e is packet error probability

P_l is packet loss probability

P_u is system unavailability probability

The packet scheduling priority weights in equation (20) were determined through simulation-based optimization using the MIMIC-III clinical database. The optimization objective was to minimize false alarms while maximizing critical event detection accuracy across diverse patient scenarios, formulated as a constrained optimization problem with $w_u+ w_r+w_l=1$

The final optimized weights are:

w_uu=0.45 (urgency priority)

$w_r r$=0.35 (reliability requirement)

w_l=0.20 (latency sensitivity)

Sensitivity analysis confirmed robust performance with less than 2% accuracy degradation under ±10% weight variations. For different clinical contexts, weights are adjusted: ICU patients use w_u=0.60 for maximum urgency response, while home monitoring emphasizes reliability with w_r=0.50."

Minor Comments

Most of the references are "touching" the previous text. Add a space between text and references. For instance: "...clinical settings[1],[2]." should be "... clinical settings [1], [2]."

Response: We have corrected all reference formatting issues throughout the manuscript. All references now include proper

spacing between text and citation numbers. Examples of corrections made:

"clinical settings [1], [2]" ✓ (was "clinical settings [1],[2]")

"remote healthcare solutions [2], [3]" ✓ (was "solutions [2],[3]")

"existing communication networks [4], [5]" ✓ (was "networks [4],[5]")

"particularly poor connectivity [6], [7]" ✓ (was "connectivity [6],[7]")

This formatting has been standardized throughout the entire manuscript for consistency.

Figure 2 should be closer to where it is referred to on the previous page.

Response: We have repositioned Figure 2 to appear immediately after its first reference in the text. The figure now appears directly following the paragraph that introduces the system architecture components, improving readability and flow.

References

1. Bharadwaj S. Peer review of "Real-Time Health Monitoring Using 5G Networks: Deep Learning–Based Architecture for Remote Patient Care". JMIRx Med 2025;6:e83423. [doi: 10.2196/83423]
2. Batool I. Real-time health monitoring using 5G networks: deep learning–based architecture for remote patient care. JMIRx Med 2025;6:e70906. [doi: 10.2196/70906]
3. Gonzalez-Canete FJ. Peer review of "Real-Time Health Monitoring Using 5G Networks: Deep Learning–Based Architecture for Remote Patient Care". JMIRx Med 2025;6:e83424. [doi: 10.2196/83424]

Edited by A Grover; submitted 03.09.25; this is a non–peer-reviewed article;accepted 03.09.25; published 01.10.25.

Please cite as:
Batool I
Author's Response to Peer Reviews of "Real-Time Health Monitoring Using 5G Networks: Deep Learning–Based Architecture for Remote Patient Care"
JMIRx Med 2025;6:e83473
URL: https://xmed.jmir.org/2025/1/e83473
doi:10.2196/83473

Authors' Response to Peer Reviews of "Real-World Performance of COVID-19 Antigen Tests: Predictive Modeling and Laboratory-Based Validation"

Miguel Bosch[1,2], PhD; Dawlyn Garcia[2], MSc; Lindsey Rudtner[2], BSc; Nol Salcedo[2], MSc; Raul Colmenares[1], MSc; Sina Hoche[2], PhD; Jose Arocha[2], MSc; Daniella Hall[2], PhD; Adriana Moreno[1], MSc; Irene Bosch[2], PhD

[1]Info Analytics Innovations, Houston, TX, United States
[2]IDX20 Inc, 166 Clinton Rd, Brookline, MA, United States

Corresponding Author:
Irene Bosch, PhD
IDX20 Inc, 166 Clinton Rd, Brookline, MA, United States

Related Articles:

Companion article: https://www.medrxiv.org/content/10.1101/2024.10.21.24315762v1

Companion article: https://med.jmirx.org/2025/1/e83476

Companion article: https://med.jmirx.org/2025/1/e83479

Companion article: https://med.jmirx.org/2025/1/e68376

(JMIRx Med 2025;6:e83474) doi:10.2196/83474

KEYWORDS

COVID-19, antigen test clinical performance; real-world data; limit of detection lateral flow test; probability of positive agreement; logistic regression

This is the authors' response to peer-review reports for "Real-World Performance of COVID-19 Antigen Tests: Predictive Modeling and Laboratory-Based Validation."

Round 1 Review

We thank the reviewers for the thoughtful feedback on our paper [1]. Below, we address each of their points.

Reviewer BH [2]

Minor Comments

1. It would be good to include a schematic/analysis/methods and an image of how the lateral flow assays look and how test band intensities are obtained. Was that done with ImageJ?

Response: We agree with the suggestion. The request has been addressed by including Figure 1B in the revised manuscript. The figure legend has also been updated and matches the new Figures 1A and 1B.

The software used is like ImageJ, but we did not use ImageJ. ImageJ was used in prior work [3]. The methodology we present here used a newly made software. The Python and R scripts that were utilized in this software have been posted on a website and are available to the public. The new References section includes the website.

2. An interesting aspect of the paper is the comparison between trained eye versus user of lateral flow assays (Figure 5). I think that adding a paragraph about the conclusions from that figure might be good in the Discussion section.

Response: The new version of the manuscript includes a paragraph in the Discussion section that explains the finding of trained eye (study staff) versus community users. Another paragraph was added to the Methods section explaining the training given to all community participants to properly report the test data.

Reviewer FZ [4]

Major Comments

1. The authors' clinical conclusions based on their prediction theory are overly optimistic.

Response: We agree to tone down our optimism. We modified the Abstract following this concern. We included in the new version of the manuscript cautionary notes and listed points of consideration.

2. The authors can explore actual clinical evaluations to determine the robustness of their prediction modeling.

Response: We agree with the follow-up plan suggested. We have addressed this concern by submitting a separate manuscript, currently "in press" at *JMIR Bioinformatics and Biotechnology*.

The preprint of the work that includes the clinical evaluation of multiple test brands is titled “Improving Antigen Test Sensitivity Estimation through Target Distribution Balancing” and currently available here [5].

3. Thus, the paper merits publication, providing the limitations are more clearly described and the conclusions are limited to the mathematical results for which the authors have proven their claims theoretically. Extension to clinical applicability is a different story yet to be told.

Response: We agree with this comment; as explained in a previous response, we have extended the clinical applicability of these methods, and the data are now in press in a *Journal of Medical Internet Research* sister journal as mentioned before.

4. The authors should be encouraged to move forward in view of the need and the poor performance of COVID-19 rapid antigen tests during the pandemic because of low sensitivity, a lack of deep understanding, and the “prevalence boundary,” a measure of when the rate of false omissions becomes too high and false negatives spread disease.

Response: The updated reference list indicates we agree with the concern. We introduced two references to illustrate that there are mitigation strategies for less sensitive diagnostic tests via serial testing (ie, testing on consecutive days during the acute COVID-19 disease stage results in an increase in the test sensitivity).

Minor Comments

5. Needs English grammar review. This could be achieved by using an artificial intelligence editor.

Response: We thank the reviewer for the suggestion. We used a grammar artificial intelligence corrector, and we introduced several changes to the original manuscript as a result of this review. A version of the modified manuscript that includes all changes to English grammar errors is available.

References

1. Bosch M, Garcia D, Rudtner L, et al. Real-world performance of COVID-19 antigen tests: predictive modeling and laboratory-based validation. JMIRx Med 2025;6:e68376. [doi: 10.2196/68376]
2. de Puig H. Peer review of “Real-World Performance of COVID-19 Antigen Tests: Predictive Modeling and Laboratory-Based Validation". JMIRx Med 2025;6:e83476. [doi: 10.2196/83476]
3. Bosch I, de Puig H, Hiley M, et al. Rapid antigen tests for dengue virus serotypes and Zika virus in patient serum. Sci Transl Med 2017 Sep 27;9(409):eaan1589. [doi: 10.1126/scitranslmed.aan1589] [Medline: 28954927]
4. Kost G. Peer review of “Real-World Performance of COVID-19 Antigen Tests: Predictive Modeling and Laboratory-Based Validation". JMIRx Med 2025;6:e83479. [doi: 10.2196/83479]
5. Bosch M, Colmenares R, Moreno A, et al. Improving real-world antigen test sensitivity estimation through target distribution balancing. MedRxiv. Preprint posted online on Oct 28, 2024. [doi: 10.1101/2024.10.25.24316137v1]

Edited by F Wu; submitted 03.09.25; this is a non–peer-reviewed article;accepted 03.09.25; published 06.10.25.

Please cite as:
Bosch M, Garcia D, Rudtner L, Salcedo N, Colmenares R, Hoche S, Arocha J, Hall D, Moreno A, Bosch I
Authors’ Response to Peer Reviews of “Real-World Performance of COVID-19 Antigen Tests: Predictive Modeling and Laboratory-Based Validation”
JMIRx Med 2025;6:e83474
URL: https://xmed.jmir.org/2025/1/e83474
doi:10.2196/83474

Authors' Response to Peer Reviews of "Effects of Interventions for the Prevention and Management of Maternal Anemia in the Advent of the COVID-19 Pandemic: Systematic Review and Meta-Analysis"

John Kyalo Muthuka[1,2], DIP PHARM, HND, BSc, PGD, MPH, PhD; Dianna Kageni Mbari-Fondo[3], PM, BEd, MSc, MPH; Francis Muchiri Wambura[4], HND, BSc, Dip-Med Lab, MPH; Kelly Oluoch[4], BPharm, MPharm, MBA, PhD; Japheth Mativo Nzioki[5], BSc (EVH), BSc (ENSc), CPH, MPH, PhD; Everlyn Musangi Nyamai[4], BScN, MPH, PhD; Rosemary Nabaweesi[6], MPH, MBChB, DrPH

[1]Department of Community Health and Health Promotion, Faculty of Public Health, Kenya Medical Training College, Mbagathi Way, Kenyatta National Hospital, Nairobi, Kenya

[2]Epidemiology/Public Health Section, KEMRI Graduate School of Health Sciences, Kenya Medical Research Institute, Nairobi, Kenya

[3]Alberta Health Services, Edmonton, AB, Canada

[4]Kenya Medical Training College, Nairobi, Kenya

[5]School of Nursing, Andrews University, Berrien Springs, MI, United States

[6]School of Global Health, Meharry Medical College, Nashville, TN, United States

Corresponding Author:
John Kyalo Muthuka, DIP PHARM, HND, BSc, PGD, MPH, PhD
Department of Community Health and Health Promotion, Faculty of Public Health, Kenya Medical Training College, Mbagathi Way, Kenyatta National Hospital, Nairobi, Kenya

Related Articles:

Companion article: https://www.medrxiv.org/content/10.1101/2024.02.18.24302492v1

Companion article: https://med.jmirx.org/2025/1/e81700

Companion article: https://med.jmirx.org/2025/1/e81699

Companion article: https://med.jmirx.org/2025/1/e82836

Companion article: https://med.jmirx.org/2025/1/e57626

(JMIRx Med 2025;6:e81711) doi:10.2196/81711

KEYWORDS

maternal anemia; anemia in pregnancy; COVID-19; pregnancy complications; meta-analysis; maternal and child health; anemia prevention; reproductive health

This is the authors' response to peer-review reports for "Effects of Interventions for the Prevention and Management of Maternal Anemia in the Advent of the COVID-19 Pandemic: Systematic Review and Meta-Analysis."

Round 1 Review

Anonymous [1]

1. Retrospective studies: The majority of the studies included in the meta-analysis [2] were retrospective epidemiological studies, which may have limitations in terms of bias, data accuracy, and causality compared to prospective studies or randomized controlled trials. This could affect the reliability and generalizability of the findings.

Response: These were the only available studies at the time. However, subgroup analysis was conducted to address these limitations and enhance the validity of the findings and conclusions.

2. Heterogeneity: The high heterogeneity identified in the pooled effect estimates suggests variability in study designs, interventions, and outcomes across the included studies. This heterogeneity can impact the interpretation of the results and the ability to draw consistent conclusions.

Response: As highlighted by other reviewers, this issue has been addressed comprehensively.

3. Publication bias: The presence of publication bias indicated by the asymmetrical funnel plot could introduce bias in the pooled effect estimates. This bias may be due to the selective reporting of studies with significant results, potentially skewing the overall findings.

Response: This has been acknowledged as a limitation and reported in the study along with measures that were implemented to mitigate it.

4. Limited scope: Some studies may not have clearly defined the age range of participants or the specific stage of the gestation period analyzed. A lack of detailed information on these aspects could limit the applicability and generalizability of the results to specific subgroups of pregnant women.

Response: While this limitation was noted, the study primarily focused on the types of interventions used during COVID-19 and how they impacted the norm in managing maternal anemia. The scope was broadly centered on reproductive age but not stratified into specific age groups.

5. Indirect effects of COVID-19: While the study focused on the direct impact of COVID-19 on maternal anemia interventions, indirect contributions of the pandemic on anemic conditions may not have been fully elucidated. Understanding these indirect effects could provide a more comprehensive view of the challenges faced during the pandemic.

Response: Subgroup and sensitivity analysis were conducted to explore the perceived indirect effects. The variability in studies likely included indirect effects, which have been explained as a potential limitation.

6. Effectiveness trends: The decreasing trend in the effectiveness of interventions against maternal anemia from 2020 to 2022 raises questions about the sustainability and adaptability of intervention strategies, especially in the context of global health emergencies. Further research is needed to explore the reasons behind this trend and potential strategies for improvement.

Response: Further research is recommended to investigate the reasons behind this trend and to develop strategies for improvement, as suggested in the Feasible Policy Recommendations section.

Reviewer IG [3]

Major Comments

1. Scientific rigor and novelty.

Strength: The focus on maternal anemia interventions during COVID-19 is unique and addresses a significant gap in the literature.

Issue: The study does not establish the novelty of its findings clearly. It cites several similar meta-analyses but does not differentiate its contribution.

Recommendation: Clarify how this meta-analysis advances existing knowledge. Are there new methodologies, expanded datasets, or novel insights?

Response: This meta-analysis advances existing knowledge through rigorous methodologies and expanded datasets from 11 articles with 6129 participants. It reveals new insights, including a 39% utility in preventing and managing maternal anemia, with significant impacts from education (28%), medicinal administration (19%), iron supplementation (17%), and intravenous ferric carboxylmaltose (15%). It highlights regional differences, particularly higher effectiveness in Africa, and underscores the need for multicenter studies and ongoing research.

2. Study design and methodology.

Inclusion criteria: The inclusion of preprints and unpublished data raises concerns about the reliability and quality of the evidence.

Suggestion: Clearly discuss the rationale for including preprints and outline strategies to mitigate biases.

Subgroup analysis: While subgroup analyses are insightful, the interpretation of heterogeneity (I^2>90% in multiple cases) is not adequately addressed. The sensitivity analyses seem to mitigate this but are not discussed in sufficient depth.

Suggestion: Incorporate a robust discussion of the potential sources of heterogeneity and its implications for the results.

Response: Inclusion criteria: While it was initially planned, preprints ultimately were not included, resolving this issue.

Subgroup analysis: Several factors contributed to this heterogeneity. Retrospective epidemiological studies dominated, with only 4 randomized controlled trials providing more robust evidence. Variability in participant age ranges and gestation stages, coupled with COVID-19's indirect effects on hemoglobin levels, contributed to this limitation. These issues are explained in detail, emphasizing the need for future research on pandemics and disasters.

3. Data presentation.

Tables and figures: Tables and figures are overly complex and lack clarity.

Suggestion: Simplify forest and funnel plots for better readability. Ensure that all figures are annotated clearly.

Forest plots: Some rate ratio confidence intervals (eg, in subgroup analysis) overlap with no-effect lines, which undermines conclusions about statistical significance.

Suggestion: Address these overlaps explicitly in the Discussion.

Response: Tables and figures: Figures have been appropriately annotated and simplified to improve readability, as generated by the software.

Forest plots: This issue has been addressed in the Discussion, aligning with the statistical outputs.

4. Statistical analysis.

Publication bias: The funnel plots indicate substantial publication bias. This is acknowledged but inadequately addressed in the Discussion.

Suggestion: Include a deeper discussion of how this bias impacts the reliability of pooled estimates.

Fixed- versus random-effects models: The rationale for choosing fixed- or random-effects models for different analyses is not well-articulated.

Suggestion: Explain this choice clearly, especially in the context of high heterogeneity.

Response: Publication bias: Publication bias likely exaggerated intervention effectiveness, potentially skewing results and conclusions. Nevertheless, comprehensive searches, statistical adjustments, and inclusion of the most relevant studies at the time ensured accurate and reliable meta-analysis outcomes, enhancing study validity. Fixed- versus random-effects models: Both models were used to balance within-study and between-study variability. This dual approach strengthened the robustness and credibility of the findings, ensuring accurate pooled estimates for maternal anemia interventions.

5. Interpretation of results.

The interpretation of intervention effects (eg, a 17% improvement for iron supplementation) does not account for clinical significance, which may differ from statistical significance.

Suggestion: Discuss the practical implications of these findings, especially in low-resource settings.

Response: The findings highlight that interventions like iron supplementation, education, and medicinal administration are critical for improving maternal anemia outcomes, particularly in low-resource settings. Even modest improvements significantly benefit maternal and infant health by reducing complications during pregnancy and childbirth.

6. Language and readability.

The manuscript is riddled with grammatical errors, unclear phrasing, and redundancies. For instance:

"The effect on prevention, control, management and or treatment of anemia was calculated and compared between the intervention and the comparator arms."

Suggestion: Simplify and clarify language to improve readability.

Acronyms (eg, RR, CI, IFA) are used without clear explanation.

Suggestion: Ensure all acronyms are defined upon first use.

Response: Grammar, phrasing, and redundancies have been thoroughly refined to enhance readability throughout the document. All acronyms have been defined upon their first use.

7. Ethical considerations.

The manuscript mentions that some data are unpublished. It is unclear whether these studies adhered to ethical guidelines.

Suggestion: Add a section on ethical considerations, particularly around the inclusion of unpublished studies.

Response: Only studies published in English between December 2019 and August 2022 were included, ensuring ethical considerations were met.

8. Discussion and Conclusion.

Weakness: The Discussion is repetitive and does not critically engage with the limitations of the study or the broader implications of the findings.

Suggestion: Provide a more focused discussion of limitations (eg, high heterogeneity, reliance on observational studies), implications for practice and policy, and recommendations for future research.

Minor Comments

1. Abstract.

Issue: The abstract lacks precision and overuses vague terms (eg, "several anemia interventions").

Suggestion: Summarize key findings clearly, avoiding overgeneralizations.

Response: Key findings have been summarized clearly, avoiding generalizations. The introduction has been streamlined to focus on the problem, knowledge gaps, and study objectives.

2. Introduction.

The Introduction is overly lengthy and includes redundant information (eg, definitions of anemia repeated multiple times).

Suggestion: Streamline the Introduction to focus on the problem, the gap in knowledge, and the study's objectives.

3. References.

References are incomplete and inconsistently formatted.

Suggestion: Ensure all references follow a standardized format (eg, APA, AMA).

Response: References have been corrected and formatted in the Vancouver style.

4. Figures.

Figures are not numbered or titled appropriately.

Suggestion: Include clear figure numbers, titles, and legends for all figures.

Response: All figures have been numbered, titled, and provided with legends for clarity.

Recommendations for Authors

Based on the above assessment, this manuscript requires major revisions. Key issues include addressing heterogeneity and publication bias in statistical analysis, improving clarity and rigor in data presentation, and enhancing language and readability.

Response: These revisions have been implemented throughout the document.

Reviewer JS [4]

Major Comments

1. There is no conclusion in this manuscript. Add a Conclusion section that summarizes the content of this study.

Response: A Conclusion section has been added at the end of the write-up, summarizing the study's key content and findings.

Minor Comments

2. Introduction section. Briefly explain the types of interventions implemented during COVID-19 to prevent anemia in pregnant women. Provide a brief explanation of the differences in anemia prevention interventions before and after COVID-19.

Response: The introduction has been updated to include this information: "During the COVID-19 pandemic, interventions adapted to include telemedicine, remote consultations, and increased community health worker involvement to address health care disruptions [5-8]. These measures aimed to ensure continued support for pregnant women [9,10]. Interventions to prevent anemia in pregnant women included iron and folic acid supplementation, dietary modifications, education and awareness programs, telemedicine, and remote consultations, as well as community-based interventions [11,12]."

3. Discussion, at the end of the Discussion section. Include the main findings of this study and emphasize their significance in addressing anemia-related challenges. Highlight the contribution of the study results to public health, particularly how the findings can inform or improve anemia prevention and treatment strategies in health care systems. Provide practical recommendations or actionable steps based on the study's outcomes that can be implemented in maternal health care policies and programs.

Response: The main findings have been incorporated into the Discussion. The meta-analysis advances knowledge by using rigorous methodologies and expanded datasets from 11 articles involving 6129 participants. It reveals new insights, including a 39% utility in preventing and managing maternal anemia. The analysis highlights the effectiveness of education (28%), medicinal administration (19%), iron supplementation (17%), and intravenous ferric carboxylmaltose (15%). Additionally, regional differences, with higher effectiveness noted in Africa, underscore the importance of multicenter studies and ongoing research.

Public health: The findings demonstrate that unforeseen pandemics may compromise anemia control, affecting interventions for maternal anemia. It is essential to screen pregnant women to identify the best intervention options for achieving optimal outcomes during crises. The impact of the COVID-19 pandemic on anemia interventions should be further studied. Key health care stakeholders must address risks posed to maternal anemia management outcomes. Future research should explore mechanisms that drive or reduce the risks of compromised maternal anemia interventions.

References

1. Anonymous. Peer review of "Effects of Interventions for the Prevention and Management of Maternal Anemia in the Advent of the COVID-19 Pandemic: Systematic Review and Meta-Analysis". JMIRx Med 2025;6:e82836. [doi: 10.2196/82836]
2. Muthuka JK, Mbari-Fondo DK, Wambura FM, et al. Effects of interventions for the prevention and management of maternal anemia in the advent of the COVID-19 pandemic: systematic review and meta-analysis. JMIRx Med 2025;6:e57626. [doi: 10.1101/2024.02.18.24302492]
3. Kumareswaran S. Peer review of "Effects of Interventions for the Prevention and Management of Maternal Anemia in the Advent of the COVID-19 Pandemic: Systematic Review and Meta-Analysis". JMIRx Med 2025;6:e81699. [doi: 10.2196/81699]
4. Winata IGS. Peer review of "Effects of Interventions for the Prevention and Management of Maternal Anemia in the Advent of the COVID-19 Pandemic: Systematic Review and Meta-Analysis". JMIRx Med 2025;6:e81700. [doi: 10.2196/81700]
5. Maddock J, Parsons S, Di Gessa G, et al. Inequalities in healthcare disruptions during the COVID-19 pandemic: evidence from 12 UK population-based longitudinal studies. BMJ Open 2022 Oct 13;12(10):e064981. [doi: 10.1136/bmjopen-2022-064981] [Medline: 36229151]
6. Yang X, Yu Y, Xu J, et al. Clinical course and outcomes of critically ill patients with SARS-CoV-2 pneumonia in Wuhan, China: a single-centered, retrospective, observational study. Lancet Respir Med 2020 May;8(5):475-481. [doi: 10.1016/S2213-2600(20)30079-5] [Medline: 32105632]
7. Singh SS, Singh LB. Training community health workers for the COVID-19 response, India. Bull World Health Organ 2022 Feb 1;100(2):108-114. [doi: 10.2471/BLT.21.286902] [Medline: 35125535]
8. Tan SY, Foo CD, Verma M, et al. Mitigating the impacts of the COVID-19 pandemic on vulnerable populations: lessons for improving health and social equity. Soc Sci Med 2023 Jul;328:116007. [doi: 10.1016/j.socscimed.2023.116007] [Medline: 37279639]
9. da Silva Lopes K, Yamaji N, Rahman MO, et al. Nutrition-specific interventions for preventing and controlling anaemia throughout the life cycle: an overview of systematic reviews. Cochrane Database Syst Rev 2021 Sep 26;9(9):CD013092. [doi: 10.1002/14651858.CD013092.pub2] [Medline: 34564844]
10. Perelman SI, Shander A, Mabry C, Ferraris VA. Preoperative anemia management in the coronavirus disease (COVID-19) era. JTCVS Open 2021 Mar;5:85-94. [doi: 10.1016/j.xjon.2020.12.020] [Medline: 34173552]

11. e-Library of Evidence for Nutrition Actions (eLENA). Exclusive breastfeeding for optimal growth, development and health of infants. World Health Organization. 2023. URL: https://www.who.int/tools/elena/interventions/exclusive-breastfeeding [accessed 2025-09-08]
12. Jin Q, Shimizu M, Sugiura M, et al. Effectiveness of non-pharmacological interventions to prevent anemia in pregnant women: a quantitative systematic review protocol. JBI Evid Synth 2024 Jun 1;22(6):1122-1128. [doi: 10.11124/JBIES-23-00081] [Medline: 38084098]

Edited by E Meinert, T Leung; submitted 01.08.25; this is a non–peer-reviewed article;accepted 01.08.25; published 06.10.25.

Please cite as:
Muthuka JK, Mbari-Fondo DK, Wambura FM, Oluoch K, Nzioki JM, Nyamai EM, Nabaweesi R
Authors' Response to Peer Reviews of "Effects of Interventions for the Prevention and Management of Maternal Anemia in the Advent of the COVID-19 Pandemic: Systematic Review and Meta-Analysis"
JMIRx Med 2025;6:e81711
URL: https://xmed.jmir.org/2025/1/e81711
doi:10.2196/81711

Authors' Response to the Peer Review of "Estimating Variance of Log Standardized Incidence Ratios Assessing Health Care Providers' Performance: Comparative Analysis Using Bayesian, Bootstrap, and Delta Method Approaches"

Solomon Woldeyohannes[1,2], BSc, MPH, PhD; Yomei Jones[1], Dipl; Paul Lawton[1], MBBS, FRACP, PhD

[1]Menzies School of Health Research, Charles Darwin University, Northern Territory, Darwin, Casuarina, Australia

[2]School of Veterinary Sciences, University of Queensland, Gatton, Australia

Corresponding Author:
Solomon Woldeyohannes, BSc, MPH, PhD
Menzies School of Health Research, Charles Darwin University, Northern Territory, Darwin, Casuarina, Australia

Related Articles:

Companion article: https://www.medrxiv.org/content/10.1101/2025.04.22.25326183v1

Companion article: https://med.jmirx.org/2025/1/e83798

Companion article: https://med.jmirx.org/2025/1/e77415

(JMIRx Med 2025;6:e83796) doi:10.2196/83796

KEYWORDS

standardized incidence ratio; SIR; performance; health care provider; machine learning; equity

This is the authors' response to the peer-review report of "Estimating Variance of Log Standardized Incidence Ratios Assessing Health Care Providers' Performance: Comparative Analysis Using Bayesian, Bootstrap, and Delta Method Approaches."

Round 1 Review

Reviewer MT [1]

General Comments

This paper [2] compares 3 approaches for estimating the variance of the log standardized incidence ratio when profiling kidney replacement therapy centers in Australia: (1) the analytical delta method, (2) the nonparametric bootstrap method (5000 resamples), and (3) Bayesian Markov chain Monte Carlo (25,500 iterations, 3 chains). Using 2005 - 2023 patient-level data from the Australia and New Zealand Dialysis and Transplant Registry and a random-effects logistic model, the authors evaluated bias, variance, and mean squared error (MSE) and visualized performance via funnel plots. Results indicated similar bias across methods but substantially lower variance and MSE for the Markov chain Monte Carlo method (bias≈0.019; variance=0.00005; MSE=0.00042) compared with the bootstrap method (variance=0.00027; MSE=0.00094).

The topic is practical and timely, yet the manuscript needs clearer model specification, interval coverage evaluation, and streamlined writing before it reaches publishable quality.

Response: We sincerely appreciate Reviewer MT's invaluable comments. We have substantially improved the content of the manuscript in response to the comments and questions of Reviewer MT. Also, thank you for acknowledging the practicality and timeliness of the presentation of our work. Herein, we have addressed point-by-point the concerns raised by Reviewer MT

Specific Comments

Major Comments

1. There are some concerns around model clarity (Poisson versus logistic language mixed; appendix is missing). Provide complete model equations, a covariate list, and software/code links and justify using the Bernoulli model for a ratio outcome.

Response:

(a) Poisson versus logistic language mixed.

As a limitation of our study, we tried to indicate that other models—for instance, the use of the Poisson model for aggregated count data (provided we aggregated the data per center and per covariate for categorical predictors)—could affect the results of the variance estimates using the 3 methods. Hence, we have replaced logistic versus Poisson with the following statement: "First, while understanding the differences between variance estimation methods is crucial for assessing the reliability of standardized incidence ratio (SIR) estimates across different centers, we did not consider how model choice

influences variance estimates and hence the resulting statistical inference. That is, we only used the hierarchical logistic regression model for modeling the binary individual-level outcome. Therefore, we did not explore the implication of using the Poisson model for aggregated data on the resulting variance estimates using the 3 methods." Please see the Limitations section for this update. Given this limitation, we are planning future work that will compare Poisson and logistic models and determine the implications on variance estimates of using the 3 methods.

(b) Appendix is missing.

Thank you! We have now included the appendices and their citations. Please see the updated manuscript for the changes we have made.

(c) Provide complete model equations.

Below are the changes that we have made based on your suggestions for the model formula:

The model equation is specified as status ~ gender+agegp+indigenous+lung+diabetes+cvd+late+bmi30+mmm+timegp+(1|centre_id). For details of the update following your suggestion, please see the Methods section of the manuscript. We have provided the model equation. Details of the covariates included in the model equation are also described in the main manuscript.

As can be seen from the above equation, we have included the following covariates: age group, gender, Indigenous status, lung disease, diabetes, obesity, cardiovascular diseases (CVD), referral status, remoteness, and time period, which are defined as follows in our dictionary.

Variable code categories and explanations: gender=a binary variable with male versus female categoriesagegp=a categorical variable with 7 categories: ≥16 - 26, ≥26 - 36, ≥36 - 46, ≥46 - 56, ≥56 - 66, ≥66 - 76 and ≥76 years indigenous=Indigenous status with Indigenous versus non-Indigenous categorieslung=lung diseases (yes vs no)

diabetes=diabetes status (yes versus no)late=late referral status (yes versus no)bmi30=binary BMI status (BMI <30 vs BMI ≥30)mmm=Monash modified model remoteness scale with 7 categories (metropolitan areas [MM1], regional centers [MM2], large rural towns [MM3], medium rural towns [MM4], small rural towns [MM5], remote communities [MM6], and very remote communities [MM7])

timegp=time periods: 2012 - 2015, 2016 - 2019, and 2020-2023

(e) Software/code link.

We would be happy to include the R codes upon acceptance of our manuscript.

(f) Justify using the Bernoulli model for a ratio outcome.

(d) Covariate list.

- gender=a binary variable with male versus female categories
- agegp=a categorical variable with 7 categories: ≥16 - 26, ≥26 - 36, ≥36 - 46, ≥46 - 56, ≥56 - 66, ≥66 - 76 and ≥76 years
- indigenous=Indigenous status with Indigenous versus non-Indigenous categories
- lung=lung diseases (yes vs no)
- diabetes=diabetes status (yes versus no)
- late=late referral status (yes versus no)
- bmi30=binary BMI status (BMI <30 vs BMI ≥30)
- mmm=Monash modified model remoteness scale with 7 categories (metropolitan areas [MM1], regional centers [MM2], large rural towns [MM3], medium rural towns [MM4], small rural towns [MM5], remote communities [MM6], and very remote communities [MM7])
- timegp=time periods: 2012 - 2015, 2016 - 2019, and 2020-2023

As can be seen from the above equation, we have included the following covariates: age group, gender, Indigenous status, lung disease, diabetes, obesity, cardiovascular diseases (CVD), referral status, remoteness, and time period, which are defined as follows in our dictionary.

Variable code categories and explanations: gender=a binary variable with male versus female categoriesagegp=a categorical variable with 7 categories: ≥16 - 26, ≥26 - 36, ≥36 - 46, ≥46 - 56, ≥56 - 66, ≥66 - 76 and ≥76 years indigenous=Indigenous status with Indigenous versus non-Indigenous categorieslung=lung diseases (yes vs no)diabetes=diabetes status (yes versus no)late=late referral status (yes versus no)bmi30=binary BMI status (BMI <30 vs BMI ≥30)mmm=Monash modified model remoteness scale with 7 categories (metropolitan areas [MM1], regional centers [MM2], large rural towns [MM3], medium rural towns [MM4], small rural towns [MM5], remote communities [MM6], and very remote communities [MM7])

timegp=time periods: 2012 - 2015, 2016 - 2019, and 2020-2023

(e) Software/code link.

We would be happy to include the R codes upon acceptance of our manuscript.

(f) Justify using the Bernoulli model for a ratio outcome.

(f) Justify using the Bernoulli model for a ratio outcome.

We have included the following 2 paragraphs in the main manuscript to justify the Bernoulli model. Please see the Model Specification and Likelihood Definition section. "This approach is methodologically valid and commonly used in Bayesian hierarchical modeling and disease mapping, especially when individual-level data are available but aggregate counts are not directly observed. Modeling binary outcomes using Bernoulli likelihoods (ie, logistic regression) is appropriate for estimating probabilities of outcomes conditional on covariates. These estimated probabilities can then be summed within groups to yield expected counts for computing the SIR [standardized incidence ratio] or relative risks. This technique allows the derivation of the SIR from model-based expected counts, which is consistent with the definition of indirect standardization [3-7]."

Application studies using this approach include Kasza et al (2013) [8] and Normand et al (2007) [9]. The application of random intercept multilevel logistic regression models for

indirectly standardizing performance measures was explored by Clark and Moore (2011) [10] using National Trauma Data Bank data for the admission year 2008. Zang et al (2013) [11] explored hierarchical logistic regression modeling under various conditions by applying Bayesian and frequentist methods.

Further details of the model specification can be seen in the appendices.

2. Interval coverage and type I error absent. Action: add a simulation or internal bootstrap to report 95% interval coverage and false-positive rates for each method.

Response: Thanks for the insight! We have now included 95% interval coverage for the bootstrapping and a credible interval for the Bayesian Markov chain Monte Carlo simulation results. Please see Table 2 for the update. We have also included abbreviations at the bottom of the table. We have included an additional table (Table 3) for the 95% false-positive rates.

3. Missing data handling unexplained. Action: quantify missingness, describe any imputation, and list all risk-adjustment covariates.

Response: We have added the following description of how the data and missingness are handled. For details, please see the Data Source and Management section. Accordingly, we have included the following two paragraphs:

"We received n=55,856 patient data on the course of treatments from February 14, 1992, to December 31, 2023. Since our initial study period was defined from January 1, 2005, to December 31, 2023, we excluded patient data from before January 1, 2005. This resulted in N=46,160 observations on the course of treatment and comorbidities data. With the revised study period definition (January 1, 2012, to December 31, 2023), following consultation with a team of chief investigators, a total of 11,586 observations were excluded (N=44,270). Due to 1743 missing observations for late referral, 808 on weight, and 188 on height variables, N=41,531 patient data were retained.

In addition, for comparison purposes, centers were split into Indigenous and non-Indigenous centers. Some centers had fewer than 20 Indigenous patients. This required considering the adequate count of Indigenous patients per center for running the hierarchical logistic regression. Accordingly, centers with fewer than 20 Indigenous patients were excluded, which resulted in n=16,243 (25,288 observations deleted) individual-level data. Moreover, we dropped patients with missing postcodes (2640 observations deleted), so that a total of N=13,603 remained. Finally, among the 13,603 observations, 3309 observations had censored status and were therefore excluded. In addition, we excluded 55 missing observations on lung diseases, cardiovascular diseases, and diabetes combined. Therefore, a total of 10,195 observations were included in our study."

In addition, the Australia and New Zealand Dialysis and Transplant Registry indicated that data submission is complete though voluntary.

We have discussed this in the Data Source and Management subsection of the Methods section and added the two references listed here to support the quality of the data:

1. McDonald SP. Australia and New Zealand Dialysis and Transplant Registry. Kidney Int Suppl (2011). 2015 May 29;5(1):39 - 44. doi: 10.1038/kisup.2015.8. PMID: 26097784.

2. ANZDATA Registry. 38th Report, Chapter 12: Indigenous People and End Stage Kidney Disease. Australia and New Zealand Dialysis and Transplant Registry. 2016. URL: https://www.anzdata.org.au/wp-content/uploads/2023/01/12_anzdata_indigenous_v3.0_160128.pdf Accessed September 23, 2025.

Minor Comments

For Table 1, add units and align decimals.

Response: We have updated the Results section by adding the "natural log scale" at the bottom of each table. Also, we have changed the decimals to 3 digits for Table 2 and Table 3. We have left the decimals in Table 1 as is to show small differences among the methods on the variability per center.

References

1. Oluwagbade E. Peer review of "Estimating Variance of Log Standardized Incidence Ratios Assessing Health Care Providers' Performance: Comparative Analysis Using Bayesian, Bootstrap, and Delta Method Approaches". JMIRx Med 2025;6:e83798. [doi: 10.2196/83798]
2. Woldeyohannes S, Jones Y, Lawton P. Estimating variance of log standardized incidence ratios assessing health care providers' performance: comparative analysis using Bayesian, bootstrap, and delta method approaches. JMIRx Med 2025;6:e77415. [doi: 10.2196/77415]
3. Hosmer DW, Lemeshow S. Confidence interval estimates of an index of quality performance based on logistic regression models. Stat Med 1995 Oct 15;14(19):2161-2172. [doi: 10.1002/sim.4780141909] [Medline: 8552894]
4. Greenland S. Modeling and variable selection in epidemiologic analysis. Am J Public Health 1989 Mar;79(3):340-349. [doi: 10.2105/ajph.79.3.340] [Medline: 2916724]
5. Goldstein H, Spiegelhalter DJ. League tables and their limitations: statistical issues in comparisons of institutional performance. J R Stat Soc Ser A Stat Soc 1996;159(3):385. [doi: 10.2307/2983325]
6. Gelman A, Hill J. Data Analysis Using Regression and Multilevel/Hierarchical Models: Cambridge University Press; 2007.
7. Lawson AB. Bayesian Disease Mapping: Hierarchical Modeling in Spatial Epidemiology, 3rd edition 2018.
8. Kasza J, Moran JL, Solomon PJ, ANZICS-Australian New Zealand Intensive Care Society Centre for Outcome and Resource Evaluation-CORE. Evaluating the performance of Australian and New Zealand intensive care units in 2009 and 2010. Stat Med 2013 Sep 20;32(21):3720-3736. [doi: 10.1002/sim.5779] [Medline: 23526209]

9. Normand SLT, Shahian DM, Krumholz HM. Statistical and Clinical Aspects of Hospital Outcomes Profiling. Statist Sci 2007;22(2):206-226. [doi: 10.1214/088342307000000096]
10. Clark DE, Moore L. Multilevel modeling. In: Li G, Baker S, editors. Injury Research: Springer; 2011. [doi: 10.1007/978-1-4614-1599-2_23]
11. Yang X, Peng B, Chen R, et al. Statistical profiling methods with hierarchical logistic regression for healthcare providers with binary outcomes. J Appl Stat 2014 Jan 2;41(1):46-59. [doi: 10.1080/02664763.2013.830086]

Abbreviations

CVD: cardiovascular disease
MSE: mean squared error

Edited by S Tungjitviboonkun; submitted 08.09.25; this is a non–peer-reviewed article;accepted 08.09.25; published 09.10.25.

Please cite as:
Woldeyohannes S, Jones Y, Lawton P
Authors' Response to the Peer Review of "Estimating Variance of Log Standardized Incidence Ratios Assessing Health Care Providers' Performance: Comparative Analysis Using Bayesian, Bootstrap, and Delta Method Approaches"
JMIRx Med 2025;6:e83796
URL: https://xmed.jmir.org/2025/1/e83796
doi:10.2196/83796

Author's Response to Peer Reviews of "Development of a Conversational Artificial Intelligence–Based Web Application for Medical Consultations: Prototype Study"

Jorge Guerra Pires, BSc, MSci, PhD

IdeaCoding Lab, Rua Timbopeba, 24, Ouro Preto, Brazil

Corresponding Author:
Jorge Guerra Pires, BSc, MSci, PhD
IdeaCoding Lab, Rua Timbopeba, 24, Ouro Preto, Brazil

Related Articles:

Companion article: https://www.medrxiv.org/content/10.1101/2023.12.31.23300681v1

Companion article: https://med.jmirx.org/2025/1/e83217

Companion article: https://med.jmirx.org/2025/1/e84443

Companion article: https://med.jmirx.org/2025/1/e56090

(JMIRx Med 2025;6:e83417) doi:10.2196/83417

KEYWORDS

artificial intelligence; ChatGPT; chatbots; conversational agent; machine learning

This is the author's response to peer-review reports for "Development of a Conversational Artificial Intelligence–Based Web Application for Medical Consultations: Prototype Study."

Round 1 Review

Anonymous [1]

Specific Comments

Major Comments

1. Please strongly consider rechecking the grammar for the paper [2].

Response: Sure, done! I have checked several times and fixed typos. Please accept my apologies for those typos.

2. Because the experience is integrated into a chatbot, indicating that providing an interactive user experience was a goal of this work, it would be good to also conduct some user research to assess the usability of the system and participants' impressions of it.

Response: I totally agree, but things are not so simple. I have actually tried to find local medical doctors to collaborate, or a local hospital, with no success. I have been working with these models for almost a decade now, yet I have never had the opportunity to work directly with medical doctors, except when they were already researchers themselves. To test the app, in addition to ethical and bureaucracy issues (and there are many), I need to find medical doctors with patients and who are friendly with those tools. According to my experience, except for medical doctors that are researchers, they do not perceive these tools well; a lack of information is one reason, a culture that is still immature is another. Some seem to see them as competitors (the models against medical doctors) and overlook the models since they make mistakes (Daniel Kahneman stressed that in his last book). This is something that has already been discussed in the literature. I have added a new section with those comments in the new version. I hope that helps. I totally agree, and I would love to collaborate with any interested researchers and with medical doctors and patients interested in testing the system.

3. It is important to also include a section discussing the potential dangers and ethical implications of deploying such a chatbot in the real world given the sensitive context and its critical implications.

Response: This is not a straightforward discussion, but I have added a section on that. This section is best done in papers dedicated to the topic. During my PhD, I was working with white box models, and this discussion is also present in this area of applied mathematics. I have added one section in the Discussion section.

Anonymous [3]

Author's note: I am almost certain this reviewer did not receive the attachments. I have spent almost one month going back and forth with the support team to make the attachments available alongside the main paper. The paper was online without the attachment. I have decided to neglect answering this reviewer's questions when it is evident the attachment would have solved

the issue, but see my replies back. The information is in the attached material.

1. I was really liking the idea of this paper and read it with great interest, but perhaps I misunderstood—I was hoping it was a chatbot that would actually give me a diagnosis (eg, once I input an image of my retina or have some conversation with it) rather than just referring me to a specialist (which would be appropriate in some cases). Please correct my understanding if I am wrong.

Response: Yes, the system is a chatbot to which you can upload images as well as information like glucose levels. The system works both with images and physiological information. The chatbot uses large language models to extract the information from text or read images with an image description. Referring the user to a specialist is a last-stage trick, a bonus. I should stress: this is a prototype, a test of concept. Thus, the system should be seen as such, not as a production-ready chatbot. Therefore, the first impression of the reviewer was correct, though I am not sure he/she read the paper completely, since the paper does what he/she thought. It is curious why he/she got this wrong impression; maybe just read the conclusions or sections of the paper.

Would it be possible this reviewer did not receive the attachments? I have spent almost one month with the support team to fix this issue; the paper was uploaded for open review without the attachments. Finally, they updated the paper with attachments, almost one month later. My fear was a review of this type; the examples of the chatbot are in the attachments. This would explain the comments of the reviewer.

2. From the initial paper idea, I got the feeling that it was going to be an app where I could start uploading images of X-rays and the chatbot, using its models, would start to tell me something about the image; instead, it seems from the example figures that all it is doing is telling the user that this is an X-ray image and to contact an expert.

Response: Yes, that is correct! Again, it is possible the reviewer did not receive the attachment, where there are examples of how the chatbot behaves.

There are several examples in the attachment; it makes a diagnosis and also can make conversation, give extra information, and more. One limitation is that I have not focused on making an open conversation chatbot; it is easy to do so, I just need to create a memory for the chatbot. I have already done that for other projects, and it works just fine.

2. Perhaps I read the paper in haste and am lacking understanding.

Response: It is also possible the reviewer did not read the paper with care and attention.

2. I would suggest showcasing a full conversation from each of your areas (X-rays, diabetes, etc), with a full screen capture of the conversation, showing an image uploaded and ending with a diagnosis (if indeed that is what your bot is capable of).

Response: Again, the reviewer may not have received the attachment. I have shown one conversation per case, one by one, with extra discussions.

I should say one thing. The app is functional, but after I announced it on Product Hunt [4], it got interest fast. This means that I had to pay OpenAI for each interaction. I even had an application programming interface (API) key that leaked. I had to put the app offline since I pay those usage fees on my own. The idea is not a production software, but rather a proof of concept. I wanted to show it works. If the reviewer lets me know when they plan to use the app, I can make it available for one week, fully functional. If the reviewer sends me an API key of their own, it can stay online as long as they want.

Here is a review from Fibaly Group:

> *Hey there, my friend! Talk about making doctor visits a little more fun and less intimidating. I'm not sure how they do it, but kudos to @ideacodinglab for creating such a clever and unique product! Keep up the awesome work!*

3. Figures 2-5 do not really give me any picture of what is going on, they just reaffirm what I thought; that is, that the bot is not actually giving any information except recognizing what type of image it is, then referring the user to a consultant? Is that correct? You really need to put some nice figures of your full flow and architecture, not too complex, but the ones you show do not really, in my opinion, provide the reader with any real value.

Response: We should keep in mind that the bot is using an API from OpenAI, which is doing most of the work as a large language model. Not sure it is my job to describe their API as they have documentation. Not sure what more I can add to the diagrams already created. The bot is a front end–like system. It just unites different techniques, that is, specialized models with the latest OpenAI API releases. The details of each model are either from OpenAI or from the user that created the specialized models. Thus, it makes no sense for me to add details of what each model is doing. I have added details in the supplementary materials for the models I have trained.

3. As a reader, I want to see what it is you have done, and as a technical person who wants to replicate your work, I would want to see your architecture in diagram form, as well as a proper flowchart of some sort (again, no need to be complex, but to a high standard as you normally see in leading journals) outlining exactly what the flow is; this should correspond to the actual app screen captures so readers can see exactly what your app does.

Response: I have already provided several diagrams. Giving further details would make it more like a tutorial, rather than an original research paper. I was asked to shorten the paper; the requested change would certainly increase the word count to more than 10,000 words.

4. Having worked on and researched chatbots, I read this with great interest but, as per my comments, I am a little confused, as it seems this bot simply refers the user to a person after recognizing an image as an X-ray, for example. I was under

the impression from the content or was half expecting the ability to input an image, be that of an X-ray or retina, and it would start giving me some diagnostic information or the like.

Response: Yes, it does. See the examples in the attachment or let me know when you are planning to use the app. I can make it available for testing for one week, with my own key, at my own cost. I could create a user section, but it will take a while to do that.

Round 2 Review

Anonymous [1]

General Comments

Thank you to the author for reading our comments and revising the paper. Many of our previous comments have been addressed; however, I still believe the writing style, grammar, and language of the paper need significant work before this can be published.

Response: Thank you for your detailed and constructive feedback. I understand the concern regarding the language and style. For this revised version, I have taken additional steps to improve clarity, grammar, and overall readability while still not changing the paper too much.

I have produced a version using large language models, which are very good at writing. I still hold that the current version is scientifically well-written, but if the editorial team decides that the text needs more proofreading, I would like to ask about the possibility of using large language models as they are free and very good at what they do. I use the latest version of ChatGPT (GPT-4). I have been using it for a while for my manuscripts, both in English and Portuguese, and it does an excellent job.

As I am applying for a full article processing charge waiver and cannot afford professional copyediting services, I have leveraged advanced language models to assist with improving the manuscript, and it would be possible to do this again. These tools have significantly evolved and can now provide high-quality formal writing support, which I have also successfully applied to previous publications. I believe the current version reflects a substantial improvement and hope it meets the journal's standards.

Anonymous [3]

General Comments

The author has given a response to each point and I am satisfied.

Response: Thank you. I appreciate the positive feedback. I am glad the revised version addressed your concerns and met your expectations.

References

1. Anonymous. Peer review of "Development of a Conversational Artificial Intelligence–Based Web Application for Medical Consultations: Prototype Study". JMIRx Med 2025:e84443. [doi: 10.2196/84443]
2. Pires JG. Development of a Conversational Artificial Intelligence–Based Web Application for Medical Consultations: Prototype Study. JMIRx Med 2025:e56090. [doi: 10.2196/56090]
3. Anonymous. Peer review of "Development of a Conversational Artificial Intelligence–Based Web Application for Medical Consultations: Prototype Study". JMIRx Med 2025;6:e83217. [doi: 10.2196/83217]
4. Robodoc. Product Hunt. URL: https://www.producthunt.com/products/robodoc [accessed 2025-09-29]

Abbreviations

API: application programming interface

Edited by T Leung; submitted 02.09.25; this is a non–peer-reviewed article;accepted 02.09.25; published 15.10.25.

Please cite as:
Pires JG
Author's Response to Peer Reviews of "Development of a Conversational Artificial Intelligence–Based Web Application for Medical Consultations: Prototype Study"
JMIRx Med 2025;6:e83417
URL: https://xmed.jmir.org/2025/1/e83417
doi:10.2196/83417

Authors' Response to Peer Reviews of "Impact of the COVID-19 Pandemic on Routine Childhood Vaccination Coverage in Ecuador From 2019 to 2021: Comparative Analysis"

Jose Sanchez[1*], MSc, MD; Alejandro Arjuna Rodriguez Sr[2*]; Kimberlly Pamela Montenegro Cuello Sr[2]

[1]Faculty of Health Sciences and Human Well-being, Universidad Indoamérica, Avenida Machala y Sabanilla, La Pradera, Quito, Ecuador

[2]Faculty of Health Sciences "Eugenio Espejo", Universidad UTE, Quito, Ecuador

[*]these authors contributed equally

Corresponding Author:
Jose Sanchez, MSc, MD
Faculty of Health Sciences and Human Well-being, Universidad Indoamérica, Avenida Machala y Sabanilla, La Pradera, Quito, Ecuador

Related Articles:

Companion article: https://www.medrxiv.org/content/10.1101/2025.03.26.25324742v1

Companion article: https://med.jmirx.org/2025/1/e84847

Companion article: https://med.jmirx.org/2025/1/e84848

Companion article: https://med.jmirx.org/2025/1/e84849

Companion article: https://med.jmirx.org/2025/1/e75293

(JMIRx Med 2025;6:e84851) doi:10.2196/84851

KEYWORDS

COVID-19 pandemic; vaccination coverage; Ecuador; immunization; routine vaccination; health disparities; vaccine hesitancy

This is the authors' response to peer review reports related to "Impact of the COVID-19 Pandemic on Routine Childhood Vaccination Coverage in Ecuador From 2019 to 2021: Comparative Analysis."

Round 1 Review

Reviewer G [1]

Major Comments

1. Please complete the manuscript [2] by adding the results and interpretation of the Joinpoint regression analyses. The authors claimed that Joinpoint regression analyses were conducted but did not present and discuss the results. More importantly, please note that the Joinpoint analysis requires at least 7 time points. The authors only included 3 time points (2019, 2020, 2021). I suggest either calculating vaccination coverage percentages for at least 7 years to run the Joinpoint analysis or just presenting the descriptive statistics for each year without using the Joinpoint analysis.

Response: We have removed all references to Joinpoint regression analysis from the manuscript. We acknowledge that our study has only 3 time points (2019, 2020, 2021), which is insufficient for Joinpoint analysis. The Methods section now clearly states that we used descriptive statistics and comparative analysis appropriate for our data structure.

2. There is a figure that plots the vaccination coverage rates in 2019, 2020, and 2021, but the authors did not provide any description or interpretation of the figure.

Response: We have added comprehensive descriptions for all figures. For example: "Figure 1 illustrates the temporal trends in vaccination coverage for key vaccines from 2019 to 2021. The visualization clearly demonstrates the progressive decline in coverage rates, with the most dramatic decreases occurring between 2020 and 2021."

3. Please add descriptions for all tables and figures.

Response: Complete descriptions have been added for both tables and all figures, explaining their content and relevance to the study findings.

4. Please be more specific in the Data Analysis section; for example, please clearly mention what was meant by trend analysis and comparative analysis and include any specific descriptive summaries and/or statistical tests you used.

Response: The data analysis section has been expanded to specify: "We calculated absolute and relative changes in

coverage between time periods" and "Coverage data were plotted over time to visualize trends and identify patterns of decline or recovery across different vaccines and regions using the *matplotlib* and *seaborn* libraries in Python."

5. Please improve the organization of the Results section. For example, the regional disparities were discussed at the end of the section, yet they were presented in the first table.

Response: The Results section has been reorganized to present (1) overall vaccination coverage trends, (2) vaccine-specific coverage analysis, and (3) regional and provincial disparities, maintaining logical flow and consistency with table presentation.

6. Please narrow the focus of the manuscript. It seems that the authors aim to characterize the changes in routine childhood vaccination before and after the COVID pandemic, but in the manuscript, the authors also mention the disparities in getting the COVID-19 vaccine among the entire Ecuador population. These seem like relatively separate topics and could possibly be studied in two manuscripts.

Response: We have removed all references to COVID-19 vaccination coverage in the general population and focused exclusively on routine childhood vaccination, as suggested. The manuscript now maintains a clear, unified focus.

7. Please support all claims with data or citations. For example, if the authors decide to also study the disparities in COVID-19 vaccine access, please include relevant data analysis results in the manuscript.

Minor Comments

8. At the start of the Data Analysis section, please cite the specific software used.

Response: We added the following: "Statistical analyses were performed using SPSS (version 28.0; IBM Corp). Trend visualization was performed using the *matplotlib* and *seaborn* libraries in Python."

9. I was wondering if there is data from after the pandemic (2022 and beyond), so the authors can examine whether routine childhood vaccination coverage went back up or kept declining.

Response: We have added the following to the Limitations section: "The analysis is limited to 2019‐2021, preventing assessment of recovery efforts that may have begun in 2022‐2023." We noted this as an important area for future research.

10. *Please clarify what Table 1 presents and why it is included.*

Response: We added an explanation: "Table 1 presents population data across Ecuador's 4 main regions and 24 provinces, providing context for understanding vaccination disparities and calculating coverage rates."

Reviewer L [3]

Major Comments

1. Clarity on methodology: The study uses observational comparative analysis and descriptive statistics but would benefit from additional details on the specific statistical tests used (eg, Joinpoint regression) and any confidence intervals or measures of significance included.

Response: We have expanded the Methods section to clarify: "We calculated absolute and relative changes in coverage between time periods using appropriate descriptive statistics. Coverage rates were calculated following World Health Organization guidelines for vaccination coverage assessment."

2. Policy and programmatic implications: While the discussion clearly outlines the negative impact on vaccination coverage, the manuscript could be strengthened by offering more specific recommendations for public health policy, especially regarding catch-up campaigns or digital infrastructure improvements to track immunization.

Response: We have significantly expanded the Policy Recommendations section, including targeted catch-up vaccination campaigns, health system strengthening, community engagement strategies, digital health innovations, and integrated service delivery models.

3. Sociodemographic context: The analysis highlights disparities but could be improved by integrating more granular sociodemographic information (eg, income, ethnicity, rurality) to provide a deeper understanding of inequities in coverage and guide targeted interventions.

Response: We have enhanced the discussion of disparities and added to the Limitations: "Detailed individual-level socioeconomic data were not available, limiting the ability to fully analyze equity impacts." We also expanded the discussion of rural/urban and Indigenous population impacts.

Minor Comments

4. Language and style: The manuscript would benefit from light editing to improve flow and reduce minor typographical and grammatical errors.

Response: The entire manuscript has been thoroughly edited for language, flow, and typographical errors.

5. Figure/table integration: Tables are rich in data, but would be more useful if the text referenced key figures and included short interpretation notes to help readers navigate large data points.

Response: We have improved integration between the text and tables/figures with specific references—for example, "Table 2 presents comprehensive coverage data showing this concerning trend"—and added interpretative notes throughout.

6. Redundancy in the Introduction: Some repetition in the early paragraphs could be streamlined to maintain reader engagement.

Response: We have eliminated redundancies in the Introduction and improved flow between paragraphs.

Reviewer M [4]

Major Comments

1. The tables and figures should be well-labeled and referenced.

Response: All tables and figures now have clear, descriptive titles and are properly referenced throughout the text with specific callouts and interpretations.

2. The limitations of the study are briefly mentioned.

Response: We have significantly expanded the Limitations section to include temporal scope limitations, the lack of individual-level socioeconomic data, causal attribution challenges, subnational granularity issues, and reliance on administrative data.

3. The statistical methods should be well-presented.

Response: The Data Analysis section has been expanded with specific details about the software used (SPSS 28.0), analytical approaches (descriptive statistics, comparative analysis, geographical analysis), and visualization methods (Python libraries).

Minor Comments

4. The Methods should be more explanatory.

Response: The Methods section has been substantially expanded to include detailed descriptions of the study design rationale, data source specifications, study population definitions, vaccination coverage metrics, ethical considerations, and data quality measures.

References

1. Wang Z. Peer review of “Impact of the COVID-19 Pandemic on Routine Childhood Vaccination Coverage in Ecuador From 2019 to 2021: Comparative Analysis”. JMIRx Med 2025;6:e84847. [doi: 10.2196/84847]
2. Sanchez J, Rodriguez AA, Cuello KPM. Impact of the COVID-19 Pandemic on Routine Childhood Vaccination Coverage in Ecuador From 2019 to 2021: Comparative Analysis. JMIRx Med 2025;6:e75293. [doi: 10.2196/75293]
3. Adekola A. Peer review of "Impact of the COVID-19 Pandemic on Routine Childhood Vaccination Coverage in Ecuador From 2019 to 2021: Comparative Analysis". JMIRx Med 2025;6:e84848. [doi: 10.2196/84848]
4. Mudashiru BA. Peer review of "Impact of the COVID-19 Pandemic on Routine Childhood Vaccination Coverage in Ecuador From 2019 to 2021: Comparative Analysis". JMIRx Med 2025;6:e84849. [doi: 10.2196/84849]

Edited by A Grover; submitted 25.09.25; this is a non–peer-reviewed article;accepted 25.09.25; published 17.10.25.

Please cite as:
Sanchez J, Rodriguez Sr AA, Cuello Sr KPM
Authors’ Response to Peer Reviews of “Impact of the COVID-19 Pandemic on Routine Childhood Vaccination Coverage in Ecuador From 2019 to 2021: Comparative Analysis”
JMIRx Med 2025;6:e84851
URL: https://xmed.jmir.org/2025/1/e84851
doi:10.2196/84851

Authors' Response to Peer Reviews of "Assessing the Limitations of Large Language Models in Clinical Practice Guideline–Concordant Treatment Decision-Making on Real-World Data: Retrospective Study"

Tobias Roeschl[1,2,3,4,5*], MD; Marie Hoffmann[2,4,5*], PhD; Djawid Hashemi[1,2,3,4], MD, PD; Felix Rarreck[2,5]; Nils Hinrichs[2,4,5], MSc; Tobias Daniel Trippel[1,2,4], MD, Prof Dr Med; Matthias I Gröschel[2,6], MD, PhD; Axel Unbehaun[2,5], MD, PD; Christoph Klein[2,5], MD, PD; Jörg Kempfert[2,5], MD, Prof Dr Med; Henryk Dreger[1,2], MD, Prof Dr Med; Benjamin O'Brien[2,7,8], MD, Prof Dr Med; Gerhard Hindricks[1,2], MD, Prof Dr Med; Felix Balzer[2,9], MD, PhD, Prof Dr Med; Volkmar Falk[2,4,5,10], MD, Prof Dr Med; Alexander Meyer[2,4,5,11], MD, Prof Dr Med

[1]Department of Cardiology, Angiology and Intensive Care Medicine, Deutsches Herzzentrum der Charité, Berlin, Germany

[10]Department of Health Sciences and Technology, Translational Cardiovascular Technologies, Institute of Translational Medicine, Swiss Federal Institute of Technology, Zürich, Switzerland

[11]Berlin Institute for the Foundations of Learning and Data – TU Berlin, Berlin, Germany

[2]Charité – Universitätsmedizin Berlin, corporate member of Freie Universität Berlin and Humboldt-Universität zu Berlin, Charitéplatz 1, Berlin, Germany

[3]Berlin Institute of Health at Charité – Universitätsmedizin Berlin, BIH Biomedical Innovation Academy, BIH Charité Digital Clinician Scientist Program, Berlin, Germany

[4]DZHK (German Centre for Cardiovascular Research), partner site Berlin, Berlin, Germany

[5]Department of Cardiothoracic and Vascular Surgery, Deutsches Herzzentrum der Charité (DHZC), Berlin, Germany

[6]Department of Infectious Diseases and Respiratory Medicine, Charité – Universitätsmedizin Berlin, Berlin, Germany

[7]Department of Cardiac Anesthesiology and Intensive Care Medicine, Deutsches Herzzentrum der Charité (DHZC), Berlin, Germany

[8]Department of Perioperative Medicine, St Bartholomew's Hospital and Barts Heart Centre, London, United Kingdom

[9]Charité – Universitätsmedizin Berlin, Institute of Medical Informatics, Berlin, Germany

[*]these authors contributed equally

Corresponding Author:
Marie Hoffmann, PhD
Charité – Universitätsmedizin Berlin, corporate member of Freie Universität Berlin and Humboldt-Universität zu Berlin, Charitéplatz 1, Berlin, Germany

Related Articles:

Companion article: http://preprints.jmir.org/preprint/74899

Companion article: https://med.jmirx.org/2025/1/e84175

Companion article: https://med.jmirx.org/2025/1/e84174

Companion article: https://med.jmirx.org/2025/1/e74899

(JMIRx Med 2025;6:e84173) doi:10.2196/84173

KEYWORDS

large language model; foundation model; reasoning model; treatment decision-making; aortic stenosis; clinical practice guidelines; medical data processing

This is the authors' response to peer-review reports for "Assessing the Limitations of Large Language Models in Clinical Practice Guideline–Concordant Treatment Decision-Making on Real-World Data: Retrospective Study."

Round 1 Review

Reviewer K [1]

1. To improve the discussion on bias in large language models (LLMs) for clinical decision-making, the study [2] should include the following aspects:

If LLMs are trained predominantly on Western medical literature or specific demographic groups, their recommendations may not generalize well to diverse patient populations. If the data used to fine-tune the model lack representation from certain ethnic, gender, or socioeconomic groups, the artificial intelligence may produce recommendations that are not universally applicable. Even with a diverse dataset, biases can arise due to model architecture, reinforcement learning strategies, or human-in-the-loop feedback mechanisms that shape model responses.

Response: Thank you for this thoughtful and important comment. We fully agree that the generalizability and fairness of LLMs in health care are significantly influenced by the composition of their training and fine-tuning data. As you rightly note, underrepresentation of certain ethnic, gender, or socioeconomic groups can lead to biased outputs and potentially widen existing health disparities. Indeed, we have also discovered, for example, bias toward transcatheter aortic valve implantation in our experiments, as indicated through the Frequency Bias Index in Figure 2 and Table S9. All LLMs were taken off-the-shelf without fine-tuning as the cohort size was limited by the inherently low incidence of eligible cases and the stringent requirements for high-quality, comprehensive patient data. Each case required detailed manual review and the generation of structured case summaries, which further constrained the pool of analyzable data. As a result, stratification and investigation of bias by additional features such as ethnic, gender, or socioeconomic features was not feasible. In the Limitations section, we have added that potential biases remain unaddressed.

2. What datasets were used? If real patient data were used, specify its source (eg, electronic health records, clinical trial data, or synthetic datasets). Provide the total number of cases or records used for testing the large language models. If synthetic data were generated, describe the method used to create the data. Were diverse age groups, genders, and ethnic backgrounds represented? A lack of diversity in data can affect the generalizability of results.

Response: Thank you for addressing this very important point. As described in the Methods section, we have used real clinical reports in PDF format from our hospital information system and extracted the content into text files. Either these text files (experiments RAW and RAW+) or manually drafted summaries (SUM and SUM+) from these text files had been used as input to the LLMs. No trial or synthetic data were used.

3. What datasets were used? If real patient data were used, specify its source (eg, electronic health records, clinical trial data, or synthetic datasets). Provide the total number of cases or records used for testing the large language models. If synthetic data were generated, describe the method used to create the data. Were diverse age groups, genders, and ethnic backgrounds represented? A lack of diversity in data can affect the generalizability of results.

Response: Thank you for your comment. This comment is identical to Comment #2, which we have addressed in detail above. To summarize: we used real clinical reports extracted from our hospital information system (electronic health records), and no synthetic or trial data were used. Additional details, including data source and sample characteristics, are provided in our response to Comment #2 and in the revised Methods section under “Study Population” and “Data Collection and Preprocessing.”

4. The study’s impact can be significantly enhanced by addressing the following challenges: Raw medical reports often include free-text narratives, physician notes, abbreviations, and inconsistencies, requiring advanced natural language processing techniques such as entity recognition, text normalization, and standardization. These reports may also contain irrelevant information, redundancies, or nonessential clinical details. Effective preprocessing is essential to filter out unnecessary content while preserving critical medical insights. A key consideration is how to optimize this preprocessing to mitigate these challenges efficiently.

Response: Thank you for this insightful comment. The central objective of our study was to assess model performance using the same type of raw clinical data that health care professionals routinely encounter, including free-text narratives and unstructured content. The rationale behind this approach was that, for real-world clinical implementation, it would be most beneficial if LLMs could generate guideline-concordant treatment recommendations directly from routine clinical documentation—without relying on curated or heavily preprocessed inputs. This would help avoid the considerable time and resource demands associated with manual or automated preprocessing pipelines. To explore this, we compared model performance on raw clinical reports with performance on highly preprocessed, structured synopses, as used in previous studies where frontier models have shown strong results. We simulated this optimized input scenario through manually drafted summaries (SUM and SUM+), which represent a best-case input condition. Replicating such preprocessing through automated means would require extensive quality control mechanisms and may still fall short of the accuracy and relevance achieved through expert curation.

Reviewer BI [3]

1. The format and provenance of the SUM (“case summary”) reports require clearer specification. Although the authors note these summaries were “manually generated,” it would be helpful to state whether they followed a standardized template, who exactly drafted them (eg, experienced cardiologists, research assistants), and which elements of the Heart Team protocol they distilled into each summary.

Response: Thank you for pointing this out. We agree that this aspect was not sufficiently described in the original manuscript. We have revised the Methods section under “Experiments” to clarify that the case summaries were manually created but adhered to a structured format: all patient characteristics documented in the heart team protocol were systematically addressed by either affirming, negating, or populating them with patient-specific values. An illustrative example is provided in Table S6.

2. The authors report that the original medical documents were saved as PDFs and later converted to plain text. It would be

helpful to clarify this process to avoid confusion, since LLMs accessed via chat interfaces or application programming interfaces often struggle with PDF inputs or text embedded in images, treating them differently from pure text. A brief discussion acknowledging this limitation—and explaining how PDF parsing was handled or validated—would help readers assess real-world applicability.

Response: We appreciate the reviewer's helpful comment. In every case, plain text—not PDF files—was provided as input. To clarify this point in the Methods section, we have added a description of the process: the text content of each PDF file was programmatically extracted using the Tesseract OCR software and concatenated into a single plain-text file, which was then used as input for the models for the RAW and RAW+ experiments.

3. Raw inputs (PDFs and summaries) were provided in German (except for BioGPT, which required translation to English). A comment in the Discussion about how model performance can vary by input language—perhaps citing studies that showed different results in Polish versus English—would contextualize the findings for non-English clinical settings:

- *Rosoł M, Gąsior JS, Łaba J, Korzeniewski K, Młyńczak M. Evaluation of the performance of GPT-3.5 and GPT-4 on the Polish Medical Final Examination.* Sci Rep. *2023;13(1):20512.*

Response: We appreciate the reviewer's thoughtful suggestion. We agree that input texts in languages other than English may pose an additional challenge for LLMs, as they are primarily trained on English-language literature. We have added a comment and the suggested citation in the Limitations section. The study you cited suggests that more recent GPT models may be more language-agnostic than previous generations, though it remains unclear whether this holds true for other languages and frontier models.

4. The Discussion section feels comparatively weak and could be strengthened by broader literature coverage. For instance, a brief discussion of input formats—pure text versus multimodal inputs—would be valuable, especially given the inclusion of GPT-4o, which handles images. Preliminary studies in this area include:

- *Günay et al. Comparison of emergency medicine specialist, cardiologist, and ChatGPT in electrocardiography assessment. Am J Emerg Med. 2024 Jun;80:51-60.*
- *Zeljkovic et al. Beyond text: the impact of clinical context on GPT-4's 12-lead electrocardiogram interpretation accuracy. Canadian J Cardiol. 2025 Jul;41(7):1406-1414.*

These compare electrocardiogram interpretation with and without accompanying clinical context and demonstrate the importance of textual input alongside images.

It would also be helpful to reference work showing that, despite similar hallucination tendencies, LLMs perform strongly on standardized exams, for example:

- *Gilson et al. How does ChatGPT perform on the USMLE? Implications for medical education and knowledge assessment. JMIR Med Educ. 2023 Feb 8;9:e45312.*
- *Novak et al. The pulse of artificial intelligence in cardiology: evaluating state-of-the-art LLMs for clinical cardiology. medRxiv. Preprint posted online on January 30, 2024.*

These additions could situate the findings within a broader context of multimodal and high-stakes assessment.

Response: We thank the reviewer for this valuable suggestion. We agree that the Discussion section benefits from a broader contextualization, particularly with respect to input formats and the evolving capabilities of multimodal models. At the current time, the diagnostic quality of multimodal models remains rudimentary, especially for images other than X-rays. As you suggested, we have added a paragraph to the Limitations section, where we stated that including imaging data in addition to the textual data would have most likely not led to a substantial improvement in model performance in our task–referring to the studies by Günay et al and Zeljkovic et al that you kindly mentioned.

In addition, we gladly added the references (Gilson et al, Novak et al) that you mentioned to the "Data Representation Affects LLM Performance" section of the Discussion to further strengthen our point that LLMs generally perform well when provided with concise and information-dense data but struggle with noisy and unprocessed clinical data.

5. As an exploratory aside, it would be interesting to evaluate how the newest reasoning-focused models (eg, "o3" or "o4") perform on this task. Although this is likely beyond the current scope, including a sentence to that effect in the manuscript's Limitations section could guide future research.

Response: We agree that in the fast-paced environment of LLM development, it is plausible that the newest reasoning-focused models might perform substantially better in our task than the reasoning models we used. We addressed this in the Limitations section.

6. For consistency and precision, when describing model access in the "Large Language Models" section (and elsewhere in the text), the manuscript should explicitly cite the exact supplementary tables or materials (eg, "see Table S1 for model details and context sizes") rather than referring generically to "the Supplementary."

Response: We agree that referring to specific supplementary tables and figures improves both clarity and precision. Accordingly, we have specified which supplementary tables and figures we are referring to throughout the manuscript.

7. In the Statistical Methods subsection, rather than stating that nonnormally distributed data were compared using the Mann-Whitney U test "for nonnormally distributed continuous variables," the phrasing could be tightened to "for variables departing from normality" or "for variables not following a normal distribution" to align with standard statistical terminology.

Response: We thank the reviewer for this constructive suggestion. We have revised the phrasing in the "Statistical Analysis" subsection of the Methods to align with standard statistical terminology. Specifically, we now refer to the use of

the Mann–Whitney *U* test for "variables departing from normality," as recommended.

Changes made to the manuscript on our end:

- We made minor adjustments to the affiliations on the title page to align with newly introduced in-house guidelines.
- In Table 2 and Table S6, we replaced the previously reported age ranges (used in accordance with medRxiv's data protection policy) with the actual patient ages, now presented as integer values.
- We replaced the term "non-LLM models" with "deterministic models" in the final paragraph before the Limitations section, as this terminology is more commonly used in recent literature and provides a more precise characterization.

References

1. Singh R. Peer review of "Assessing the Limitations of Large Language Models in Clinical Practice Guideline–Concordant Treatment Decision-Making on Real-World Data: Retrospective Study". JMIRx Med 2025;6:e84175. [doi: 10.2196/84175]
2. Roeschl T, Hoffmann M, Hashemi D, et al. Assessing the limitations of large language models in clinical practice guideline–concordant treatment decision-making on real-world data: retrospective study. JMIRx Med 2025;6:e84173. [doi: 10.2196/84173]
3. Novak A. Peer review of "Assessing the Limitations of Large Language Models in Clinical Practice Guideline–Concordant Treatment Decision-Making on Real-World Data: Retrospective Study". JMIRx Med 2025;6:e84174. [doi: 10.2196/84174]

Abbreviations

LLM: large language model

Edited by A Grover; submitted 15.09.25; this is a non–peer-reviewed article;accepted 15.09.25; published 03.11.25.

Please cite as:
Roeschl T, Hoffmann M, Hashemi D, Rarreck F, Hinrichs N, Trippel TD, Gröschel MI, Unbehaun A, Klein C, Kempfert J, Dreger H, O'Brien B, Hindricks G, Balzer F, Falk V, Meyer A
Authors' Response to Peer Reviews of "Assessing the Limitations of Large Language Models in Clinical Practice Guideline–Concordant Treatment Decision-Making on Real-World Data: Retrospective Study"
JMIRx Med 2025;6:e84173
URL: https://xmed.jmir.org/2025/1/e84173
doi:10.2196/84173

Authors' Response to Peer Reviews of "Evaluating the Financial Factors Influencing Maternal, Newborn, and Child Health in Africa: Tobit Regression and Data Envelopment Analysis"

Youssef Er-Rays[1]; Meriem M'dioud[2]; Hamid Ait-Lemqeddem[2]; Badreddine El Moutaqi[1]

[1]Polydisciplinary Faculty of Larache, Abdelmalek Essaadi University, Tetouan, Morocco

[2]École nationale des sciences appliquées, Ibn Tofail University, Kenitra, Morocco

Corresponding Author:
Youssef Er-Rays
Polydisciplinary Faculty of Larache, Abdelmalek Essaadi University, Tetouan, Morocco

Related Articles:

Companion article: https://www.medrxiv.org/content/10.1101/2024.02.22.24303217v1

Companion article: https://med.jmirx.org/2025/1/e85382

Companion article: https://med.jmirx.org/2025/1/e85383

Companion article: https://med.jmirx.org/2025/1/e59703

(JMIRx Med 2025;6:e85578) doi:10.2196/85578

KEYWORDS

financial determinants; maternal, newborn, and child health; health care efficiency; Africa; health expenditure; data envelopment analysis; Tobit regression

This is the authors' response to peer-review reports for "Evaluating the Financial Factors Influencing Maternal, Newborn, and Child Health in Africa: Tobit Regression and Data Envelopment Analysis."

Round 1 Review

Reviewer GP [1]

The manuscript [2] provides a valuable contribution to the understanding of health care system efficiency in Africa, particularly in the context of maternal, newborn, and child health (MNCH). The study is ethical, with appropriate use of data from reputable sources such as the World Health Organization, and the methods employed—data envelopment analysis and Tobit regression—are suitable for assessing the technical efficiency of health care systems across 46 African countries.

The material is original, and the paper addresses a significant gap in the literature by focusing on the financial and efficiency factors impacting MNCH in Africa. Related work is discussed and cited adequately, although a few more recent studies could be included to strengthen the literature review.

The writing is generally clear, though there are some areas where the discussion of the results could benefit from more detail. The study methods are appropriate for the research objectives, and the data used appear to be valid and reliable. The findings are significant and present actionable insights for policymakers, especially in terms of understanding the inefficiencies in the health care systems that impact MNCH outcomes.

The conclusions are reasonable and are supported by the data, although more detailed recommendations for practical application could enhance the paper's impact. The topic is certainly of interest to the readership, as it addresses key issues surrounding health care efficiency and the achievement of Sustainable Development Goal 3 in Africa.

Overall, I recommend the manuscript for publication with minor revisions to improve the clarity of some sections and provide more detailed policy recommendations.

Response: The authors express gratitude for the constructive comments on the manuscript, recognizing the study's contribution to understanding health care system efficiency in Africa, particularly in MNCH. The authors have clarified and expanded the discussion of the results, updated the literature review to include recent studies, and added more concrete policy recommendations in the Conclusion section. These changes aim to strengthen the theoretical framework and align the research with the existing body of work. The authors thank the reviewers for their feedback for improving the manuscript's quality and clarity.

Reviewer GW [3]

General Comments

Although the title might initially seem misleading, the paper tackles an essential issue of efficiency in delivering MNCH services in Africa. Given the well-known challenges facing maternal and newborn health in the region, the importance of this study cannot be overstated.

Specific Comments

Major Comments

1. The whole abstract is more about general efficiency in health care systems and less about the financial factors influencing MNCH in Africa. The title has to reflect what the study actually presents.

Response: Thank you for this comment. We will address it.

2. If the key aim of the study was to determine how financial factors influence MNCH, one would then expect to see, in the abstract, the extent of the influence of financial factors such as health expenditure, coverage index, and expenditure per capita.

Response: Thank you for this comment. We will address it.

3. The Introduction section starts with a presentation of the number of women dying in 2020 and the number of children dying in 2021. These data are not supported with any citations. It is also a little strange that for the number of women dying, the study refers to 2020 data, while for that of children, the study refers to 2021.

Response: We greatly appreciate your keen observation on the limitations of the data. You are absolutely right in your assessment. The primary challenge we encountered during this analysis was indeed the unavailability of comprehensive, directly reported numbers for maternal and newborn deaths specifically for the entire African continent in both 2020 and 2021.

- The Introduction section starts with a presentation of the number of women dying in 2020 and the number of children dying in 2021. These data are not supported with any citations. It is also a little strange that for the number of women dying, the study refers to 2020 data, while for that of children, the study refers to 2021.
- **Response:** We greatly appreciate your keen observation on the limitations of the data. You are absolutely right in your assessment. The primary challenge we encountered during this analysis was indeed the unavailability of comprehensive, directly reported numbers for maternal and newborn deaths specifically for the entire African continent in both 2020 and 2021.

4. The Introduction section does not provide sufficient motivation for investigating efficiency in health systems. The first paragraph presents the maternal health challenges, while the second paragraph quickly goes to the methods for establishing efficiency. There is no connection as to why investigating efficiency is necessary.

Response: Thank you for this comment. We will address it.

5. The Introduction section provides a descriptive review of other studies without depth. It lists the different studies without synthesizing them. It would be useful if they were at least lifted up to present issues/themes so it is easy to connect with what the study is about.

Response: Thank you for this comment. We will address it.

6. There is a sentence in the Methods section (Data Sources and Variables) that says “Input, output, and explanatory variables were selected to assess the accuracy of the WHO [World Health Organization]...” Are there three types of variables in your study?

Response: Thank you for this comment. Yes, we show them in Table 1.

7. What is presented as stages of data envelopment analysis does not go further to describe how the study made use of these stages. Much of the presentations are about what these stages are and sometimes the historical background. It would be useful to put more emphasis on how the study used these stages so it assures the credibility and reliability of the findings.

Response: Thank you for this comment. We will address it.

8. The presented results do not have a clear foundation from the methods. The chain of evidence from the data is lacking, from their processing to their results.

Response: Thank you for this comment. We will address it.

9. The parameters presenting the results are not clearly defined. It says “...26% with a score of 1.” There is no proper introduction of the ranges for a reader to comprehend the meaning of a score of 1. It also states that “...average efficiency score (TE-VRS) across all countries is 0.849 for VRS [variable returns to scale].”

Response: Thank you for this comment. We will address it.

10. The Discussion section needs revising. It does not directly connect to the findings of the study, despite the challenges of the results. The Discussion section further presents a couple of statistics, especially in the first and second paragraphs, without sources or a clear connection to the findings.

Response: Thank you for this comment. We will address it.

References

1. Olorunyomi TD. Peer review of “Evaluating the Financial Factors Influencing Maternal, Newborn, and Child Health in Africa: Tobit Regression and Data Envelopment Analysis”. JMIRx Med 2025;6:e85382. [doi: 10.2196/85382]
2. Er-Rays Y, M’dioud M, Ait-Lemqeddem H, El Moutaqi B. Evaluating the financial factors influencing maternal, newborn, and child health in Africa: Tobit regression and data envelopment analysis. JMIRx Med 2025;6:e59703. [doi: 10.2196/59703]

3. Mahundi M. Peer review of “Evaluating the Financial Factors Influencing Maternal, Newborn, and Child Health in Africa: Tobit Regression and Data Envelopment Analysis”. JMIRx Med 2025;6:e85383. [doi: 10.2196/85383]

Edited by F Wu; submitted 09.10.25; this is a non–peer-reviewed article;accepted 09.10.25; published 28.11.25.

Please cite as:
Er-Rays Y, M'dioud M, Ait-Lemqeddem H, El Moutaqi B
Authors' Response to Peer Reviews of "Evaluating the Financial Factors Influencing Maternal, Newborn, and Child Health in Africa: Tobit Regression and Data Envelopment Analysis"
JMIRx Med 2025;6:e85578
URL: https://xmed.jmir.org/2025/1/e85578
doi:10.2196/85578

Impact of a Point-of-Care Ultrasound Training Program on the Management of Patients With Acute Respiratory or Circulatory Failure by In-Training Emergency Department Residents (IMPULSE): Before-and-After Implementation Study

Sandra Bieler[1], MD; Stephan von Düring[2], MD; Damien Tagan[3], MD; Olivier Grosgurin[4], MD; Thierry Fumeaux[5], MD, MBA

[1]Médecin cheffe, Service des Urgences, Hôpital de Sion, Sion, Switzerland

[2]Faculté de Médecine de l'Université de Genève, Hôpitaux Universitaires de Genève, Genève, Switzerland

[3]Service des Soins critiques, Hôpital Riviera Chablais, Rennaz, Switzerland

[4]Service de médecine interne générale et Service des Urgences, Hôpitaux Universitaires de Genève, Genève, Switzerland

[5]Hirslanden Geneva Clinics, Geneva, Switzerland

Corresponding Author:
Sandra Bieler, MD
Médecin cheffe, Service des Urgences, Hôpital de Sion, Sion, Switzerland

Related Articles:

Companion article: https://www.medrxiv.org/content/10.1101/2023.09.28.23295699v1

Companion article: https://med.jmirx.org/2025/1/e72144

Companion article: https://med.jmirx.org/2025/1/e72092

Abstract

Background: Due to its diagnostic accuracy, point-of-care ultrasound (POCUS) is becoming more frequently used in the emergency department (ED), but the feasibility of its use by in-training residents and the potential clinical impact have not been assessed.

Objective: This study aimed to assess the feasibility of implementing a structured POCUS training program for in-training ED residents, as well as the clinical impact of their use of POCUS in the management of patients in the ED.

Methods: IMPULSE (Impact of a Point-of Care Ultrasound Examination) is a before-and-after implementation study evaluating the impact of a structured POCUS training program for ED residents on the management of patients admitted with acute respiratory failure (ARF) and/or circulatory failure (ACF) in a Swiss regional hospital. The training curriculum was organized into 3 steps and consisted of a web-based training course; an 8-hour, practical, hands-on session; and 10 supervised POCUS examinations. ED residents who successfully completed the curriculum participated in the postimplementation phase of the study. Outcomes were time to ED diagnosis, rate and time to correct diagnosis in the ED, time to prescribe appropriate treatment, and in-hospital mortality. Standard statistical analyses were performed using chi-square and Mann-Whitney *U* tests as appropriate, supplemented by Bayesian analysis, with a Bayes factor (BF)>3 considered significant.

Results: A total of 69 and 54 patients were included before and after implementation of the training program, respectively. The median time to ED diagnosis was 25 (IQR 15 - 60) minutes after implementation versus 30 (IQR 10 - 66) minutes before implementation, a difference that was significant in the Bayesian analysis (BF=9.6). The rate of correct diagnosis was higher after implementation (51/54, 94% vs 36/69, 52%; *P*<.001), with a significantly shorter time to correct diagnosis after implementation (25, IQR 15 - 60 min vs 43, IQR 11 - 70 min; BF=5.0). The median time to prescribe the appropriate therapy was shorter after implementation (47, IQR 25 - 101 min vs 70, IQR 20 - 120 min; BF=2.0). Finally, there was a significant difference in hospital mortality (9/69, 13% vs 3/54, 6%; BF=15.7).

Conclusions: The IMPULSE study shows that the implementation of a short, structured POCUS training program for ED residents is not only feasible but also has a significant impact on their initial evaluation of patients with ARF and/or ACF, improving diagnostic accuracy, time to correct diagnosis, and rate of prescribing the appropriate therapy and possibly decreasing

hospital mortality. These results should be replicated in other settings to provide further evidence that implementation of a short, structured POCUS training curriculum could significantly impact ED management of patients with ARF and/or ACF.

(JMIRx Med 2025;6:e53276) doi:10.2196/53276

KEYWORDS

point-of-care ultrasonography; training program; emergency department; acute respiratory failure; acute circulatory failure

Introduction

Acute respiratory failure (ARF) and acute circulatory failure (ACF) are common causes of emergency department (ED) admissions and are associated with significant morbidity, mortality, and ED resource use. Timely and appropriate management can reduce these outcomes but depends on an efficient diagnostic workup [1]. In a high proportion of EDs around the world, patients received first-line treatment by junior in-training physicians. Traditionally, the workup is guided by history taking and physical examination, which have been shown to be inaccurate in the ED, particularly when performed by less experienced physicians [2-4]. Basic laboratory and imaging tests are often supplemented with more advanced modalities, such as transthoracic echocardiography or computed tomography (CT), at the expense of increased ED length of stay, resource use, and potential adverse events [5-7]. Point-of-care ultrasound (POCUS), performed by nonradiologists or noncardiologists, is a noninvasive bedside diagnostic tool that has been shown to be highly accurate in identifying the etiologic cause of ARF or ACF, with no significant side effects [8-20]. POCUS is now included in many training programs for emergency physicians [21-27]. However, it is still unclear if the diagnostic accuracy of POCUS translates into a clinically relevant difference in patient outcomes [18,28-33]. Despite these limitations, the American College of Physicians guidelines recommend the use of POCUS in addition to standard diagnostic procedures in patients with acute dyspnea [34,35]. In most of the published studies, POCUS was performed by trained experts who were not directly responsible for the patient and were often blinded to clinical data, which does not reflect real-life conditions where patients are initially managed by junior or in-training residents.

We designed the IMPULSE (Impact of a Point-of-Care Ultrasound Examination) study to evaluate the feasibility and impact of implementing a structured POCUS training program for in-training ED residents in the first-line management of patients admitted for ACF and/or ARF. A before-and-after implementation study design was chosen to avoid the methodological problems associated with blinding and randomization in a single-center study [35].

Methods

Study Design and Intervention

IMPULSE is a single-center, before-and-after, observational, implementation study of a structured POCUS training program for ED residents (first or second year of internal medicine training) at a regional hospital (Hôpital de Nyon, Switzerland). During the preimplementation period (phase 1), patient management was unchanged, and POCUS could only be performed on demand by trained attending physicians as part of the standard ED management implemented since 2010. Only 1 in-training ED resident per 12-hour shift participated in the study.

During the intervention phase, a group of residents in training (first and second year after graduation) were enrolled in the AURUS (Association des urgentistes et réanimateurs intéressés à l'ultrasonographie) training program, organized into 3 steps and in accordance with the European Society of Intensive Care Medicine consensus document [36-38]:

- A 20-hour, web-based course on general principles of ultrasound as well as theoretical and practical aspects of image acquisition and interpretation in transthoracic, cardiac, vascular, pulmonary, and abdominal POCUS [39]: The module includes a formal assessment of knowledge through a multiple-choice questionnaire, which must be completed to proceed to the next step.
- An 8-hour, practical, hands-on session in which POCUS examinations are performed on healthy volunteers and simulators in groups of 3 students under the supervision of an instructor, focusing on the technical aspects of obtaining interpretable images: The session includes a formal assessment of image acquisition and interpretation skills. This assessment is mandatory to proceed to the next step.
- The practice of at least 10 directly supervised POCUS full examinations, performed under real conditions in the ED: This includes a formal assessment of the ability to acquire, interpret, and integrate good-quality images into clinical management.

At the end of the training process, residents who met all training objectives were enrolled in the postimplementation phase (phase 2). Similar to phase 1, only 1 ED resident per shift participated in the study. A Sparq Ultrasound System (Philips AG Healthcare) was used for all POCUS examinations, which were performed with a 4 - 12 MHz linear probe and a 1 - 4 MHz phased array probe. POCUS was requested to be performed as soon as possible on all enrolled patients, in parallel with the clinical evaluation and according to a standardized protocol evaluating 18 specific sonographic signs (Figure 1), looking for echographic signs of pulmonary embolism, left heart failure, hypovolemic state, tamponade, pneumonia, pneumothorax, or abdominal disease. All POCUS images were recorded, and a standardized case report form was completed by the resident (Figure 2). All images were mandatorily reviewed by a POCUS-trained attending physician, directly or subsequently, to confirm the findings.

All other diagnostic procedures were used at the discretion of the clinician, including a basic POCUS performed by the

attending physician and an advanced ultrasound performed by a fully trained radiologist or cardiologist.

Figure 1. Point-of-care ultrasound (POCUS) protocol evaluating specific sonographic signs: (1) internal jugular vein; (2) to (5) anterior pulmonary view or anterior axillary line view; (6) and (7) posterobasal pulmonary view; (8) inferior vena cava; (9) parasternal short- and long-axis cardiac views; (10) apical four-chamber cardiac view; (11) subcostal cardiac view; (12) hepatorenal space; (13) splenorenal space; (14) suprapubic view; and (15) to (18) femoropopliteal veins.

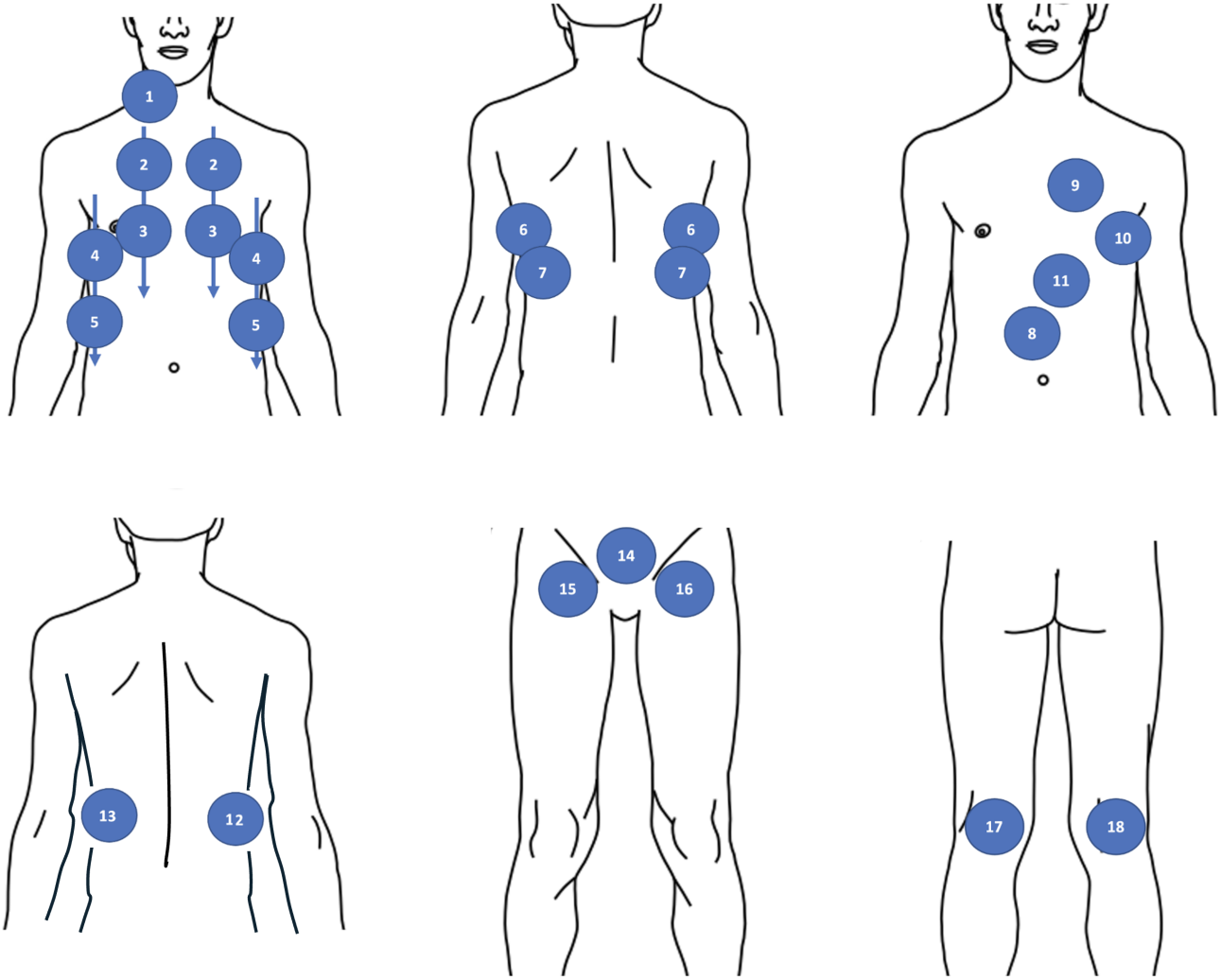


Figure 2. Case report form (adapted from the original form in French). COPD: chronic obstructive pulmonary disease; IMPULSE: Impact of a Point-of-Care Ultrasound Examination.

Case report form

Start of care : __ h__ IMPULSE identification number :

Time of diagnosis : __h__

Type of diagnosis (one or more) :

1. Pneumonia
2. Asthma/COPD exacerbation
3. Pulmonary embolism
4. Pneumothorax
5. Pericardial effusion/tamponade
6. Pleural effusion
7. Cardiac failure (acute pulmonary edema)
8. Myocardial infarction or myocarditis with cardiogenic shock
9. Septic shock
10. Gastrointestinal bleeding
11. Intraperitoneal bleeding
12. Other (specify clearly) :

Treatment prescription time : __h__

Treatment prescribed (one or more)

1. Antibiotics
2. Bronchodilators
3. Corticosteroids
4. Diuretics
5. Noninvasive ventilation (NIV)
6. Anticoagulants
7. Vasopressors
8. Coronarography
9. Abdominal surgery
10. Gastroscopy
11. Other (specify clearly ; examples : pericardial or pleural drainage, intravenous lysis, thrombectomy, arterial embolization,… .

Time of diagnosis modification (if applicable) : __h__ New diagnosis :

Comment :

Time of treatment modification (if applicable) : __h__ New treatment:

Comment :

Patient Inclusion and Exclusion Criteria

In both phases, all consecutive adult patients (aged ≥18 years) presenting with ARF and/or ACF were screened for inclusion in the study. ARF was defined by (1) the presence of either signs of respiratory distress or a respiratory rate greater than 20 breaths/min and (2) an oxygen saturation measured using pulse oximetry of <92% on room air or the need to administer oxygen to maintain a saturation of ≥92%. ACF was defined by (1) the presence of a systolic blood pressure <90 mm Hg and (2) clinical signs of hypoperfusion (agitation or altered consciousness, skin mottling, or oliguria) or hyperlactatemia (>2.0 mmol/L).

Exclusion criteria were a known or immediate diagnosis (such as ST-elevation myocardial infarction or referral for an externally determined diagnosis), the need for immediate lifesaving measures (such as cardiopulmonary resuscitation), trauma, palliative care, and patient refusal of care.

In order to preserve the organization of the ED and to favor the admission of patients for whom uninterrupted care seemed likely, the final admission of patients and the start of observation were left to the discretion of the attending physician, based on his or her assessment of the ED situation and workload.

Data Collection

On a standardized case report form, the ED resident recorded various times (start of observation, time of diagnosis, start of diagnosis-specific therapy, and end of ED stay). Diagnoses and therapies were also reported according to a specified list (Figure 1). The participating resident was equipped with an audio recorder, which was started at first contact with the patient. All recordings were kept confidential only to the investigators, who analyzed them to verify the written data reported. Based on these data, the time to diagnosis; time to prescription of targeted, appropriate treatment; and length of stay in the ED were calculated and rounded to 5-minute intervals. The hospital discharge summary was retrospectively analyzed to compare the diagnosis made during the ED stay with the final hospital diagnosis and to assess in-hospital mortality.

Statistical Analysis

All data were analyzed with the free, open-source JASP tool (University of Amsterdam). Median and IQR values are reported for descriptive statistics of continuous variables, and absolute numbers and proportions are reported for categorical variables. Differences in proportions of categorical variables between phases were analyzed by chi-square test, with a significant level set at $P<.05$. Differences in continuous variables and time intervals between phases were analyzed with a Mann-Whitney *U* test, completed by a Bayesian approach. For this analysis, the alternative hypothesis was that the time intervals would be greater in phase 1 than in phase 2, with a prior probability described by a Cauchy distribution centered around zero and with a width parameter of 1.00. This width parameter was chosen after an equivalence, Bayesian, independent-samples (2-tailed) *t* test analysis and corresponds to a probability of 50% that the effect size lies between −1.000 and 1.000. The statistical significance of the Bayesian analysis was expressed with the Bayes factor (BF), where a value between 3 and 10 is considered moderate evidence, and a value over 10 represents strong evidence. For hospital mortality comparison between the 2 phases, a Bayesian analysis was also performed, with an independent binomial analysis, with fixed rows.

Ethical Considerations

The study was approved by the regional ethics committee (Commission Cantonale d'Ethique du Canton de Vaud; protocol 194/15). Due to the observational design of the study and the fact that the practice of POCUS was already part of the usual care in the ED of the institution, a signed individual informed consent was only required for the use of the data collected for the study. Therefore, in order not to delay the management of the patients, brief verbal information was given to the patient at the beginning of the observation. Full information about the study was then given to the patient as soon as possible. Definite enrollment and data analysis were completed only after individually signed informed consent. If the patient refused to participate, then all study materials were destroyed. No compensation was provided to patients, and all data were anonymized for analysis purposes.

Results

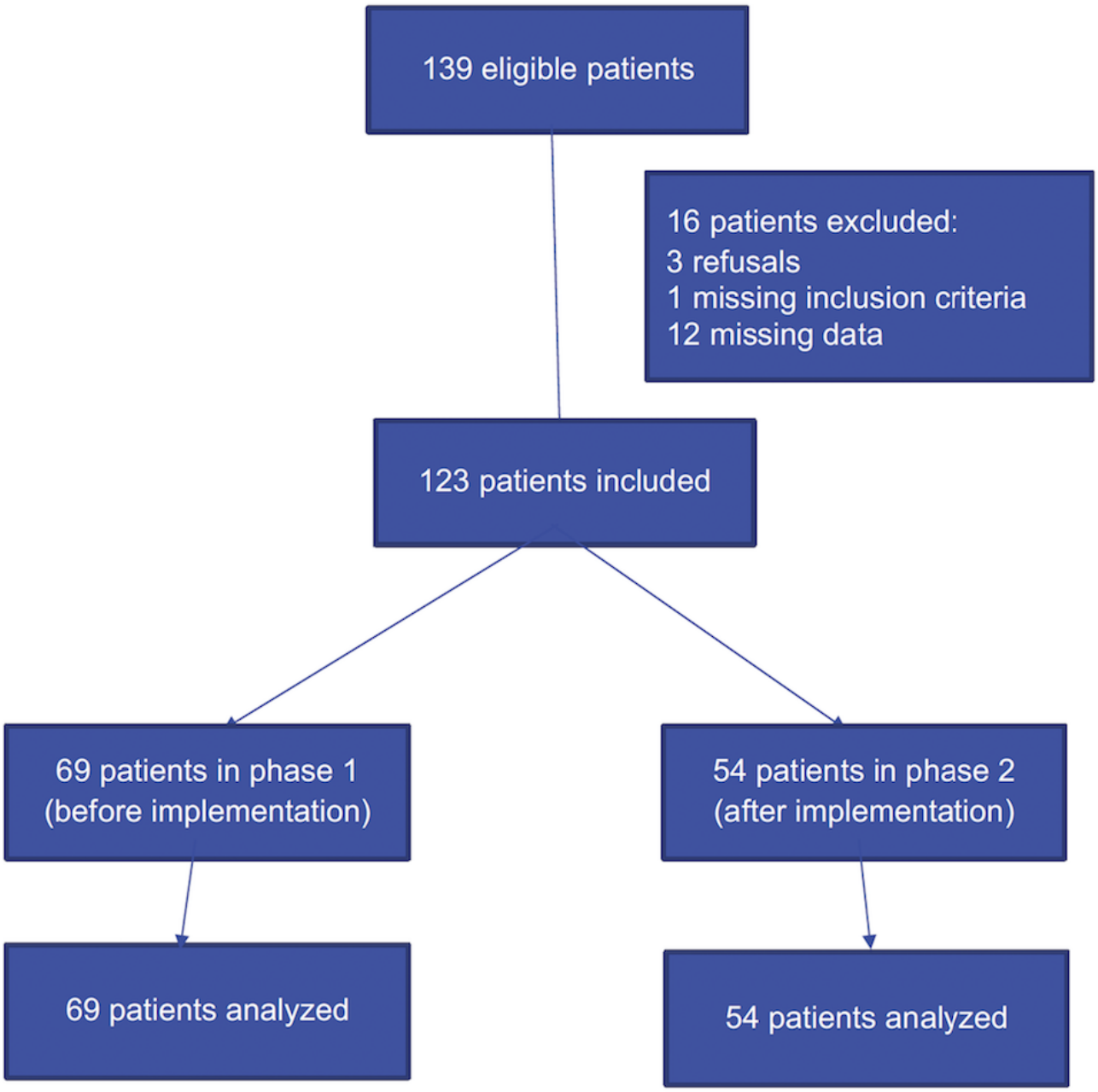
In-Training ED Residents

For ED organizational purposes, in-training residents (first or second year of training in internal medicine) were assigned to groups of 6-8 people for a 6-month rotation period. During each 12-hour shift, a resident was responsible for the first-line management of patients with ARF and/or ACF, under the supervision of an emergency medicine specialist. From September 4, 2015, to May 28, 2016 (a total of 268 days; phase 1), 14 residents participated in the observational phase, with no changes to the organization or process of usual care. Twelve interns successfully completed the AURUS training course from May 29, 2016, to September 14, 2016. Thereafter, from September 15, 2016, to February 7, 2018 (a total 511 days; phase 2), they were able to perform an immediate POCUS when managing a patient with ARF and/or ACF, which was the only difference from the observational phase 1.

Patients

During the whole study period, 139 patients were enrolled, but 3 (2.2%) patients withdrew consent to participate, 1 (0.7%) patient was excluded due to incomplete inclusion criteria, and 12 (8.6%) patients were excluded due to missing data, leaving 123 (88.5%) patients for the analysis (Figure 3). A total of 69 patients were included during phase 1 and 54 patients were included during phase 2. In the final analysis, of the 123 patients, 117 (95.1%) presented with ARF and 20 (16.3%) presented with ACF, of whom 14 (11.4%) presented with a combination of ARF (Figure 3).

Figure 3. CONSORT (Consolidated Standards of Reporting Trials) study flowchart.



The median age of the enrolled patients was 77 (IQR 70 - 84) years, and most patients were enrolled for respiratory distress (116/123, 94.3%) and hypoxemia (117/123, 95.1%). The admission characteristics of the enrolled patients are representative of the usual patients with ARF and/or ACF who present to the ED (Table 1).

Table . Patients characteristics at admission.

	Total population (n=123)	Phase 1 (n=69)	Phase 2 (n=54)
Age (years), median (IQR)	77 (70 - 84)	78 (70 - 86)	75 (70 - 82)
Female sex, n (%)	63 (51.2)	37 (53.6)	26 (48.1)
Prehospital medicalized care, n (%)	19 (15.4)	8 (11.6)	11 (20.4)
Medical history, n (%)			
COPD[a]	35 (28.5)	21 (30.4)	14 (25.9)
Asthma	9 (7.3)	5 (7.2)	4 (7.4)
Ischemic heart disease	41 (33.3)	21 (30.4)	20 (37)
Chronic heart failure	38 (30.9)	17 (24.6)	21 (38.9)
Active or past smoking	44 (35.8)	22 (31.9)	22 (40.7)
Immunosuppressive therapy	4 (3.3)	4 (5.8)	0 (0)
Pulmonary hypertension	7 (5.7)	4 (5.8)	3 (5.6)
Chronic kidney disease	44 (35.8)	22 (31.9)	22 (40.7)
Inclusion criteria, n (%)			
Respiratory distress	116 (94.3)	64 (92.8)	52 (96.3)
Hypoxemia (SpO_2[b]<92%)	117 (95.1)	66 (95.7)	51 (94.4)
Hypotension (SBP[c]<90 mm Hg)	22 (17.9)	14 (20.3)	8 (14.8)
Clinical hypoperfusion	20 (16.3)	12 (17.4)	8 (14.8)
Admission vital signs, median (IQR)			
SpO_2 (%)	89 (83 - 92)	89 (86 - 93)	88.0 (80-92)
Respiratory rate (breaths/min)	28 (24 - 32)	28 (25 - 32)	28 (24 - 34)
Heart rate (beats/min)	100 (87 - 117)	100 (88 - 115)	105 (85 - 126)
SBP (mm Hg)	132 (112 - 152)	132 (115 - 158)	130 (110 - 152)
DBP[d] (mm Hg)	76 (61 - 89)	76 (60 - 90)	75 (63 - 89)
Laboratory values, median (IQR)			
pH	7.40 (7.35 - 7.45)	7.41 (7.35 - 7.45)	7.40 (7.36 - 7.45)
pO_2[e] (kPa)	8.2 (7.1 - 9.8)	8.3 (7.4 - 10.2	7.7 (6.7 - 9.2)
pCO_2[f] (kPa)	4.9 (4.1 - 6.3)	5.0 (4.4 - 6.0)	4.8 (3.9 - 6.8)
Lactate (mmol/L)	1.75 (1.40 - 2.75)	1.80 (1.40 - 2.85)	1.70 (1.40 - 2.28)
Creatinine (µmol/L)	104 (73 - 151)	108 (73 - 152)	98 (74 - 148)
Hemoglobin (g/L)	130 (115 - 143)	130 (114 - 144)	133 (116 - 143)
BNP[g] (ng/L)	398 (185 - 924)	267 (164 - 680)	566 (311 - 1044)
D-dimers (ug/mL)	1392 (643 - 2800)	1125 (697 - 1437)	2273 (453 - 4474)
CRP[h] (mg/L)	44 (15 - 104)	43 (15 - 95)	49 (16 - 147)

[a]COPD: chronic obstructive pulmonary disease.

[b]SpO_2: oxygen saturation.

[c]SBP: systolic blood pressure.

[d]DBP: diastolic blood pressure.

[e]pO_2: partial pressure of oxygen.

[f]pCO_2: partial pressure of carbon dioxide.

[g]BNP: brain natriuretic peptide.

[h]CRP: C-reactive protein.

General ED Management

The median ED stay duration was 230 (IQR 160 - 300) minutes. During their ED stay, of the 123 patients, 98 (79.7%) had a chest x-ray, 40 (32.5%) had a chest CT scan, and 47 (38.2%) had a POCUS performed by a senior supervisor. Pneumonia was the most frequent diagnosis (n=42, 34.1%), followed by acute heart failure (n=41, 33.3%). Antibiotics (n=64, 52%) and diuretics (n=49, 39.8%) were the most frequently prescribed therapies during ED stay. Except for 2 patients (1 death and 1 home discharge), all patients were hospitalized—in half (n=58, 47.2%) of the cases, in the intensive care unit (Table 2).

Table . Emergency department (ED) management.

	Total population (n=123)	Phase 1 (n=69)	Phase 2 (n=54)
Imaging, n (%)			
Chest x-ray	98 (79.7)	65 (94.2)	33 (61.1)
Thoracic CT[a]	40 (32.5)	21 (30.4)	19 (35.2)
Abdominal CT	14 (11.4)	5 (7.2)	9 (16.7)
Abdominal ultrasound	4 (3.3)	4 (5.8)	0 (0)
Transthoracic echocardiography	3 (2.4)	2 (2.9)	1 (1.9)
POCUS[b] by senior physician	47 (38.2)	24 (34.8)	23 (42.6)
ED diagnosis, n (%)			
Pneumonia	42 (34.1)	26 (37.7)	16 (29.6)
Acute heart failure	41 (33.3)	19 (27.5)	22 (40.7)
Acute exacerbation of COPD[c]	13 (10.6)	9 (13)	4 (7.4)
Nonpulmonary sepsis	11 (8.9)	8 (11.6)	3 (5.6)
Pulmonary embolism	5 (4.1)	1 (1.4)	4 (7.4)
Pericardial effusion	3 (2.4)	0 (0)	3 (5.6)
Cardiogenic shock	2 (1.6)	1 (1.4)	1 (1.9)
Other diagnosis	6 (4.9)	5 (7.2)	1 (1.9)
Specific ED therapies, n (%)[d]			
Antibiotics	64 (52)	39 (56.5)	25 (46.3)
Diuretic therapy	49 (39.8)	24 (34.8)	25 (46.3)
Bronchodilators	27 (22)	18 (26.1)	9 (16.7)
Noninvasive ventilation	25 (20.3)	15 (21.7)	10 (18.5)
Steroids	17 (13.8)	10 (14.5)	7 (13)
Anticoagulation	14 (11.4)	5 (7.2)	9 (16.7)
Vasopressors	12 (9.8)	6 (8.7)	6 (11.1)
Patient destination after ED stay, n (%)			
Ward	58 (47.2)	36 (52.2)	22 (40.7)
ICU[e]	58 (47.2)	30 (43.5)	28 (51.9)
Other hospital (ICU or ward)	5 (4.1)	2 (2.9)	3 (5.6)
Home	1 (0.8)	1 (1.4)	0 (0)
Death in the ED	1 (0.8)	0 (0)	1 (1.9)

[a]CT: computed tomography.

[b]POCUS: point-of-care ultrasound.

[c]COPD: chronic obstructive pulmonary disease.

[d]Some patients may have received more than 1 therapy.

[e]ICU: intensive care unit.

Comparison Between Phase 1 and Phase 2

The proportion of final diagnoses retained at the end of hospitalization that confirmed the ED diagnosis was 52.2% (36/69) in phase 1 and 94.4% (51/54) in phase 2, a highly significant difference (χ^2_1=26.146, *P*<.001; Table 3).

Table . Confirmation of emergency department diagnosis during hospital diagnosis: contingency table[a].

	Diagnostic confirmed during hospital stay	
	No, n (%)	Yes, n (%)
Phase 1 (n=69)	33 (47.8)	36 (52.2)
Phase 2 (n=54)	3 (5.6)	51 (94.4)
Total (n=123)	36 (29.3)	87 (70.7)

[a]χ^2_1=26.146, *P*<.001.

Compared to phase 1, there was a statistically significant and clinically relevant decrease in the median time to final ED diagnosis in phase 2 (30, IQR 10 - 65 min vs 25, IQR 15 - 60 min; BF=9.6; Table 4).

Table . Emergency department (ED) time intervals.

	Phase 1 (n=69)	Phase 2 (n=54)	BF[a,b]	*P* value[c]
Time to final diagnosis (min), median (IQR)	30 (10 - 65)	25 (15 - 60)	9.56	.33
Time to final confirmed diagnosis (min), median (IQR)	43 (10 - 70)	25 (15 - 60)	5.02	.33
Time to administer a correct therapy (min), median (IQR)	70 (20 - 120)	47 (25 - 101)	1.96	.31
Duration of ED stay (min), median (IQR)	238 (163 - 300)	230 (160 - 275)	4.18	.42

[a]BF: Bayes factor.

[b]Alternative hypothesis: phase 1>phase 2; prior probability: Cauchy, scale 1.0.

[c]*P* value calculated with the Mann-Whitney *U* test.

When the ED diagnosis was confirmed during the hospital stay, the time to diagnosis in the ED was significantly shorter in phase 2 (25, IQR 15 - 60 min vs 43, IQR 10-70 min; BF=5.0), a difference of 18 minutes that is only moderately significant in the Bayesian analysis but clinically highly relevant. Finally, the time to order and start the most appropriate therapy was reduced from 70 (IQR 20 - 120) minutes in phase 1 to 47 (IQR 25 - 101) minutes in phase 2 (BF=2.0). There was also a reduction in the length of stay in the ED, which was significant in the Bayesian analysis, although probably not clinically relevant (Table 4).

Finally, in-hospital mortality was reduced in phase 2 (3/54, 5.6% vs 9/69, 13% in phase 1), a difference that was highly significant in Bayesian analysis (BF=16.04; Table 5).

Table . Hospital mortality: contingency table[a,b].

	Hospital mortality	
	Alive, n (%)	Dead, n (%)
Phase 1 (n=69)	60 (87)	9 (13)
Phase 2 (n=54)	51 (94.4)	3 (5.6)
Total (n=123)	111 (90.2)	12 (9.8)

[a]χ^2_1=1.93, *P*=.16.

[b]Bayesian analysis (independent multinomial analysis, with an alternate hypothesis: phase 1>phase 2): Bayes factor=16.04.

Due to the small population sample, we did not perform a formal statistical analysis of patient characteristics, components of ED management, distribution of diagnoses, and therapies administered (Tables 1 and 2). Nevertheless, we demonstrated a substantial decrease in the number of chest radiographs performed during phase 2, with an increase in the number of CT scans performed during the ED stay. In phase 1, according to the study design, a POCUS was performed by a senior attending physician in 34.8% (24/69) of the patients, whereas in phase 2, all patients had a POCUS performed by a junior attending physician, with a second POCUS performed by a senior attending physician in almost half (23/54, 42.6%) of the cases (Table 2).

Discussion

Principal Findings

The objective of the IMPULSE study was to investigate the feasibility and impact of implementing a brief, structured training program for ED residents on the management of patients admitted for ARF and/or ACF and their subsequent clinical outcomes. A before-and-after implementation design was selected to emulate the methodology of a randomized controlled trial, while mitigating the potential for contamination bias between the 2 groups. The only difference in the management of patients between the 2 phases was the immediate use of POCUS by the in-training resident in charge in the first-line treatment of the patient. The POCUS training curriculum (AURUS) was chosen for its established presence within the institution and its alignment with the updated recommendations concerning the training objectives of the current guidelines [37,38]. We hypothesized that the immediate use of POCUS by the junior physician after the short AURUS training would improve the diagnostic process, as compared by the later use by a senior physician.

The implementation of the structured, AURUS-based, POCUS program was not only associated with a significantly higher diagnostic accuracy rate but also a shorter delay of diagnosis, particularly when the ED diagnosis was later confirmed during the hospital stay. Our results also suggest that implementing a POCUS training program for in-training residents may be associated with a quicker implementation of the most appropriate therapeutic intervention, and possibly to a reduction in mortality rates, although the study design and the small sample size render the results susceptible to several potential biases. These findings align with those of a previous publication, which demonstrated that the use of POCUS by physicians of varying levels of experience was associated with an improved administration of appropriate therapies, despite no improvement in diagnostic accuracy [40]. This difference in diagnostic accuracy may be due to the more senior level of experience of the involved physicians in the published study, compared to our observation, as the diagnostic contribution of the ultrasound is probably greater for less experienced physicians.

It is also pertinent to consider some of the secondary findings of the IMPULSE study. In both phases of the study, the senior attending physician could conduct a POCUS examination; this occurred in nearly half of the cases in the postimplementation phase, a proportion that exceeds that observed in the preimplementation phase (Table 2). This may have been for verification purposes, but it is also possible that a POCUS conducted by a junior physician may prompt more experienced physicians to perform it with greater frequency, as a ripple effect. Similarly, although this finding should be interpreted with caution, there was a reduction in the number of chest x-rays performed during phase 2 (61.1% of patients only). This suggests that the POCUS may be used in place of this examination. Conversely, the number of CT scans performed during phase 2 was higher, which could be interpreted in two ways. It could be a negative effect of the POCUS, whereby supervisors performed more CT scans to confirm or reject a diagnosis made by their junior colleagues. The observed increase in the number of POCUS examinations performed by supervisors suggests that this may be a more positive effect. POCUS provides a more comprehensive assessment of the clinical situation, leading to a more appropriate use of advanced diagnostic modalities. Subsequent studies will likely address these findings and may confirm these trends, while providing clarification regarding the causes of the observed increase in CT scan use.

Our results show that the reported intervention is not only feasible but also that it has an impact on the clinical management process and possibly on the patient outcome. To the best of our knowledge, these data represent the inaugural demonstration of the clinical impact of a POCUS training program for ED residents. If replicated, they could substantiate the implementation of POCUS in conjunction with history taking and clinical examination by ED residents as a primary diagnostic tool.

Strengths and Limitations

The IMPULSE study has several notable strengths. The study design reflects the typical circumstances observed in most EDs, wherein patients are initially managed by junior physicians under the guidance of more experienced, senior medical professionals. The characteristics of the included patients and the diagnoses made in the ED demonstrate that this study sample is representative of the population of interest for the use of POCUS, with significant associated morbidity and mortality. The before-and-after study design circumvents the contamination bias observed in several previously published studies. The initial phase reflects the typical practice of most EDs, wherein POCUS is conducted by senior physicians at a relatively late stage, serving as a control for the subsequent postimplementation phase. Interestingly, the rate of inaccurate ED diagnosis during the phase 1 reflects the usual diagnostic accuracy for the management of patients who present to the ED [41-43].

The signal of a clinically relevant impact on the patient outcome is an interesting finding, as morbidity and mortality are the usual end points of choice for ED interventional studies. As POCUS is not a therapeutic procedure, the effect on outcome can only be driven by a quicker and more appropriate administration of efficient therapies. Therefore, our findings of quicker and more accurate diagnosis may explain the reduction of hospital mortality that was evidenced in our small population sample.

It is important to consider the limitations of the IMPULSE study, including the lack of randomization. However, as there is a risk of contamination between the two arms of a randomized controlled trial, we therefore elected to use a before-and-after implementation design as the optimal method to achieve quasi-randomization of patients to limit this risk. A cluster randomization of multiple centers with successive implementation would likely have been the optimal design in this situation; however, it was not feasible to organize. A second limitation is the single-center design and the limited sample of included patients, despite a lengthy recruitment period, particularly in phase 2, with 1 included patient every 9 days. This illustrates the challenges inherent in conducting

single-center studies in smaller institutions lacking dedicated clinical research resources. Notwithstanding this significant limitation, the studied population is representative of the typical patients with ARF and/or ACF admitted to the majority of EDs globally, as evidenced by their characteristics and corresponding diagnoses. It would be prudent to reproduce our results in other clinical settings, with the inclusion of a larger sample of patients, before any firm conclusion can be made regarding the impact of implementing a POCUS training program for in-training ED residents. These limitations do not affect the fundamental conclusions of the presented results.

Conclusion

In conclusion, the IMPULSE study demonstrates that a brief, structured training program for ED residents is both feasible and enables them to use POCUS as a primary tool for the initial management of patients presenting with ARF and/or ACF. The deployment of POCUS by these less experienced physicians may be associated with an increase in diagnostic accuracy, comparable to that observed in published data on POCUS use by experienced ED physicians. Furthermore, it may be associated with a reduction in the time required for in-training residents to reach a correct diagnosis and with a more rapid and appropriate prescription of a specific therapy, which may result in a decrease in hospital mortality. The results of the IMPULSE study also validate the AURUS training curriculum, demonstrating that this structured, stepwise approach to training is not only feasible but also efficient. These results must be replicated and validated in other settings with larger patient samples. However, the methodology presented herein is appropriate for limiting the issues of blinding and randomization in the study of such diagnostic tools and may be used by future studies.

Conflicts of Interest

None declared.

References

1. Ray P, Birolleau S, Lefort Y, et al. Acute respiratory failure in the elderly: etiology, emergency diagnosis and prognosis. Crit Care 2006;10(3):R82. [doi: 10.1186/cc4926] [Medline: 16723034]
2. Leuppi JD, Dieterle T, Koch G, et al. Diagnostic value of lung auscultation in an emergency room setting. Swiss Med Wkly 2005 Sep 3;135(35-36):520-524. [doi: 10.4414/smw.2005.10886] [Medline: 16323069]
3. Wipf JE, Lipsky BA, Hirschmann JV, et al. Diagnosing pneumonia by physical examination: relevant or relic? Arch Intern Med 1999 May 24;159(10):1082-1087. [doi: 10.1001/archinte.159.10.1082] [Medline: 10335685]
4. Mulrow CD, Lucey CR, Farnett LE. Discriminating causes of dyspnea through clinical examination. J Gen Intern Med 1993 Jul;8(7):383-392. [doi: 10.1007/BF02600079] [Medline: 8410400]
5. Collins SP, Lindsell CJ, Storrow AB, Abraham WT, ADHERE Scientific Advisory Committee, Investigators and Study Group. Prevalence of negative chest radiography results in the emergency department patient with decompensated heart failure. Ann Emerg Med 2006 Jan;47(1):13-18. [doi: 10.1016/j.annemergmed.2005.04.003] [Medline: 16387212]
6. Al Aseri Z. Accuracy of chest radiograph interpretation by emergency physicians. Emerg Radiol 2009 Mar;16(2):111-114. [doi: 10.1007/s10140-008-0763-9] [Medline: 18779982]
7. Brenner DJ, Hall EJ. Computed tomography--an increasing source of radiation exposure. N Engl J Med 2007 Nov 29;357(22):2277-2284. [doi: 10.1056/NEJMra072149] [Medline: 18046031]
8. Lichtenstein DA, Mezière GA. Relevance of lung ultrasound in the diagnosis of acute respiratory failure: the BLUE protocol. Chest 2008 Jul;134(1):117-125. [doi: 10.1378/chest.07-2800] [Medline: 18403664]
9. Lichtenstein D. FALLS-protocol: lung ultrasound in hemodynamic assessment of shock. Heart Lung Vessel 2013;5(3):142-147. [Medline: 24364005]
10. Lichtenstein D, Goldstein I, Mourgeon E, Cluzel P, Grenier P, Rouby JJ. Comparative diagnostic performances of auscultation, chest radiography, and lung ultrasonography in acute respiratory distress syndrome. Anesthesiology 2004 Jan;100(1):9-15. [doi: 10.1097/00000542-200401000-00006] [Medline: 14695718]
11. Copetti R, Soldati G, Copetti P. Chest sonography: a useful tool to differentiate acute cardiogenic pulmonary edema from acute respiratory distress syndrome. Cardiovasc Ultrasound 2008 Apr 29;6:16. [doi: 10.1186/1476-7120-6-16] [Medline: 18442425]
12. Agricola E, Bove T, Oppizzi M, et al. "Ultrasound comet-tail images": a marker of pulmonary edema: a comparative study with wedge pressure and extravascular lung water. Chest 2005 May;127(5):1690-1695. [doi: 10.1378/chest.127.5.1690] [Medline: 15888847]
13. Chavez MA, Shams N, Ellington LE, et al. Lung ultrasound for the diagnosis of pneumonia in adults: a systematic review and meta-analysis. Respir Res 2014 Apr 23;15(1):50. [doi: 10.1186/1465-9921-15-50] [Medline: 24758612]
14. Cibinel GA, Casoli G, Elia F, et al. Diagnostic accuracy and reproducibility of pleural and lung ultrasound in discriminating cardiogenic causes of acute dyspnea in the emergency department. Intern Emerg Med 2012 Feb;7(1):65-70. [doi: 10.1007/s11739-011-0709-1] [Medline: 22033792]

15. Volpicelli G, Lamorte A, Tullio M, et al. Point-of-care multiorgan ultrasonography for the evaluation of undifferentiated hypotension in the emergency department. Intensive Care Med 2013 Jul;39(7):1290-1298. [doi: 10.1007/s00134-013-2919-7] [Medline: 23584471]
16. Zanobetti M, Poggioni C, Pini R. Can chest ultrasonography replace standard chest radiography for evaluation of acute dyspnea in the ED? Chest 2011 May;139(5):1140-1147. [doi: 10.1378/chest.10-0435] [Medline: 20947649]
17. Xirouchaki N, Magkanas E, Vaporidi K, et al. Lung ultrasound in critically ill patients: comparison with bedside chest radiography. Intensive Care Med 2011 Sep;37(9):1488-1493. [doi: 10.1007/s00134-011-2317-y] [Medline: 21809107]
18. Laursen CB, Sloth E, Lambrechtsen J, et al. Focused sonography of the heart, lungs, and deep veins identifies missed life-threatening conditions in admitted patients with acute respiratory symptoms. Chest 2013 Dec;144(6):1868-1875. [doi: 10.1378/chest.13-0882] [Medline: 23948720]
19. Sasmaz MI, Gungor F, Guven R, Akyol KC, Kozaci N, Kesapli M. Effect of focused bedside ultrasonography in hypotensive patients on the clinical decision of emergency physicians. Emerg Med Int 2017;2017:6248687. [doi: 10.1155/2017/6248687] [Medline: 28357139]
20. Gartlehner G, Wagner G, Affengruber L, et al. Point-of-care ultrasonography in patients with acute dyspnea: an evidence report for a clinical practice guideline by the American College of Physicians. Ann Intern Med 2021 Jul;174(7):967-976. [doi: 10.7326/M20-5504] [Medline: 33900798]
21. Abbasi S, Farsi D, Hafezimoghadam P, Fathi M, Zare MA. Accuracy of emergency physician-performed ultrasound in detecting traumatic pneumothorax after a 2-h training course. Eur J Emerg Med 2013 Jun;20(3):173-177. [doi: 10.1097/MEJ.0b013e328356f754] [Medline: 22828649]
22. Bustam A, Noor Azhar M, Singh Veriah R, Arumugam K, Loch A. Performance of emergency physicians in point-of-care echocardiography following limited training. Emerg Med J 2014 May;31(5):369-373. [doi: 10.1136/emermed-2012-201789] [Medline: 23428721]
23. Jones AE, Tayal VS, Kline JA. Focused training of emergency medicine residents in goal-directed echocardiography: a prospective study. Acad Emerg Med 2003 Oct;10(10):1054-1058. [doi: 10.1111/j.1553-2712.2003.tb00574.x] [Medline: 14525737]
24. Moore CL, Rose GA, Tayal VS, Sullivan DM, Arrowood JA, Kline JA. Determination of left ventricular function by emergency physician echocardiography of hypotensive patients. Acad Emerg Med 2002 Mar;9(3):186-193. [doi: 10.1111/j.1553-2712.2002.tb00242.x] [Medline: 11874773]
25. Mandavia DP, Aragona J, Chan L, Chan D, Henderson SO. Ultrasound training for emergency physicians--a prospective study. Acad Emerg Med 2000 Sep;7(9):1008-1014. [doi: 10.1111/j.1553-2712.2000.tb02092.x] [Medline: 11043996]
26. Filopei J, Siedenburg H, Rattner P, Fukaya E, Kory P. Impact of pocket ultrasound use by internal medicine housestaff in the diagnosis of dyspnea. J Hosp Med 2014 Sep;9(9):594-597. [doi: 10.1002/jhm.2219] [Medline: 24891227]
27. Counselman FL, Sanders A, Slovis CM, Danzl D, Binder LS, Perina DG. The status of bedside ultrasonography training in emergency medicine residency programs. Acad Emerg Med 2003 Jan;10(1):37-42. [doi: 10.1111/j.1553-2712.2003.tb01974.x] [Medline: 12511313]
28. Bellone A, Etteri M, Maino C, Bonetti C, Natalizi A. The role of bedside ultrasound in the diagnosis and outcome of patients with acute respiratory failure. Emerg Care J 2013;9(1):e2. [doi: 10.4081/ecj.2013.e2]
29. Kanji HD, McCallum J, Sirounis D, MacRedmond R, Moss R, Boyd JH. Limited echocardiography-guided therapy in subacute shock is associated with change in management and improved outcomes. J Crit Care 2014 Oct;29(5):700-705. [doi: 10.1016/j.jcrc.2014.04.008] [Medline: 24857642]
30. Pirozzi C, Numis FG, Pagano A, Melillo P, Copetti R, Schiraldi F. Immediate versus delayed integrated point-of-care-ultrasonography to manage acute dyspnea in the emergency department. Crit Ultrasound J 2014;6(1):5. [doi: 10.1186/2036-7902-6-5] [Medline: 24940478]
31. Jones AE, Tayal VS, Sullivan DM, Kline JA. Randomized, controlled trial of immediate versus delayed goal-directed ultrasound to identify the cause of nontraumatic hypotension in emergency department patients. Crit Care Med 2004 Aug;32(8):1703-1708. [doi: 10.1097/01.ccm.0000133017.34137.82] [Medline: 15286547]
32. Atkinson PRT, McAuley DJ, Kendall RJ, et al. Abdominal and Cardiac Evaluation with Sonography in Shock (ACES): an approach by emergency physicians for the use of ultrasound in patients with undifferentiated hypotension. Emerg Med J 2009 Feb;26(2):87-91. [doi: 10.1136/emj.2007.056242] [Medline: 19164614]
33. Zieleskiewicz L, Lopez A, Hraiech S, et al. Bedside POCUS during ward emergencies is associated with improved diagnosis and outcome: an observational, prospective, controlled study. Crit Care 2021 Jan 22;25(1):34. [doi: 10.1186/s13054-021-03466-z] [Medline: 33482873]
34. Cid X, Canty D, Royse A, et al. Impact of point-of-care ultrasound on the hospital length of stay for internal medicine inpatients with cardiopulmonary diagnosis at admission: study protocol of a randomized controlled trial-the IMFCU-1 (Internal Medicine Focused Clinical Ultrasound) study. Trials 2020 Jan 8;21(1):53. [doi: 10.1186/s13063-019-4003-2] [Medline: 31915052]
35. Prager R, Wu K, Bachar R, et al. Blinding practices during acute point-of-care ultrasound research: the BLIND-US meta-research study. BMJ Evid Based Med 2021 Jun;26(3):110-111. [doi: 10.1136/bmjebm-2020-111577] [Medline: 33177166]

36. Tagan D, Fumeaux T, Beaulieu Y, Association des urgentistes et réanimateurs intéressés par l'ultrasonographie. Innovative concept in ultrasonography training targeted for the intensivist using e-learning and simulation [Article in French]. Rev Med Suisse 2015 Apr 1;11(468):785-786. [doi: 10.53738/REVMED.2015.11.468.0785] [Medline: 26021141]
37. Azarnoush K, Guechi Y, Schmutz T, Peyrony O, Fumeaux T, Ribordy V. Point-of-care ultrasonography, update on practices and a concept of implementation in an emergency department [Article in French]. Rev Med Suisse 2019 May 8;15(650):984-989. [doi: 10.53738/REVMED.2019.15.650.0984] [Medline: 31066531]
38. Expert Round Table on Ultrasound in ICU. International expert statement on training standards for critical care ultrasonography. Intensive Care Med 2011 Jul;37(7):1077-1083. [doi: 10.1007/s00134-011-2246-9] [Medline: 21614639]
39. Online courses: basic (whole body). POCUS Academy. URL: https://pocus.academy/en_GB/lesson-types/online/basic-level [accessed 2025-02-14]
40. Msolli MA, Sekma A, Marzouk MB, et al. Bedside lung ultrasonography by emergency department residents as an aid for identifying heart failure in patients with acute dyspnea after a 2-h training course. Ultrasound J 2021 Feb 9;13(1):5. [doi: 10.1186/s13089-021-00207-9] [Medline: 33559777]
41. Peng A, Rohacek M, Ackermann S, et al. The proportion of correct diagnoses is low in emergency patients with nonspecific complaints presenting to the emergency department. Swiss Med Wkly 2015 May;145:w14121. [doi: 10.4414/smw.2015.14121] [Medline: 25741894]
42. Sikka R, Tommaso LH, Kaucky C, Kulstad EB. Diagnosis of pneumonia in the ED has poor accuracy despite diagnostic uncertainty. Am J Emerg Med 2012 Jul;30(6):881-885. [doi: 10.1016/j.ajem.2011.06.006] [Medline: 21855251]
43. Johnson T, McNutt R, Odwazny R, Patel D, Baker S. Discrepancy between admission and discharge diagnoses as a predictor of hospital length of stay. J Hosp Med 2009 Apr;4(4):234-239. [doi: 10.1002/jhm.453] [Medline: 19388065]

Abbreviations:

ACF : acute circulatory failure
ARF : acute respiratory failure
AURUS: Association des urgentistes et réanimateurs intéressés à l'ultrasonographie
BF : Bayes factor
CT: computed tomography
ED: emergency department
IMPULSE: Impact of a Point-of-Care Ultrasound Examination
POCUS: point-of-care ultrasound

Edited by A Schwartz, E Meinert; submitted 02.10.23; peer-reviewed by Anonymous; revised version received 06.01.25; accepted 30.01.25; published 03.03.25.

Please cite as:
Bieler S, von Düring S, Tagan D, Grosgurin O, Fumeaux T
Impact of a Point-of-Care Ultrasound Training Program on the Management of Patients With Acute Respiratory or Circulatory Failure by In-Training Emergency Department Residents (IMPULSE): Before-and-After Implementation Study
JMIRx Med 2025;6:e53276
URL: https://xmed.jmir.org/2025/1/e53276
doi:10.2196/53276

Effect of Thermal and Vibration Changes on Automated External Defibrillator Circuit Boards: Finite Element Analysis Study

Saidi Olayinka Olalere, MSc

Department of Mechanical Engineering, Georgia Southern University, 1332 Southern Drive, Statesboro, GA, United States

Corresponding Author:
Saidi Olayinka Olalere, MSc
Department of Mechanical Engineering, Georgia Southern University, 1332 Southern Drive, Statesboro, GA, United States

Related Articles:

Companion article: https://arxiv.org/abs/2309.14747v1

Companion article: https://med.jmirx.org/2025/1/e80142

Companion article: https://med.jmirx.org/2025/1/e80137

Companion article: https://med.jmirx.org/2025/1/e80135

Abstract

Background: An automated external defibrillator (AED) is a device that is used to prevent sudden death by delivering an electrical shock to restore the heart rhythm when experiencing cardiac arrest.

Objective: This study was performed to analyze the vibration and thermal changes experienced by an AED medical device when exposed to shocks caused by patients' reactions, vibrations from mobile and air ambulances, and heat changes due to the battery component on the circuit board.

Methods: Basically, AED is made from plastic, with the external parts containing the display, buttons, pad socket, and speaker, while the internal part entails the circuit boards comprising components such as resistors, capacitors, inductors, and integrated circuits, among others. In this study, the AED was modeled with the Ansys Workbench 2020 and calibrated based on static and dynamic loading to verify the static displacement and determine the first set of five frequencies obtained based on the unprestressed conditions.

Results: Using the prestressed analysis with modifications, the next set of frequencies was obtained with an error margin of 0.0003% between each frequency. The modeled circuit board was used to examine the vibration and dynamic analysis for the rigid board. Similarly, thermal analysis was conducted on the modeled circuit board with the battery serving as the heat source. The rate of dissipation of heat around the board and its effect on the circuit components was evaluated.

Conclusions: The modeled circuit board was reinforced with more support structures to mitigate the deformation effect. The deformation peaked at 33.172 mm, with minimum deformation at the edges of the board. Components with greater height, such as capacitors, experienced more pronounced deformation. Therefore, it is suggested that flat capacitors of lesser height would be suitable for future designs. Additionally, significant heat dissipation from the battery suggests a need for better dissipation pathways.

(JMIRx Med 2025;6:e53208) doi:10.2196/53208

KEYWORDS

circuit board; automated external defibrillator; heart; cardiology; vibration; thermal changes; medical devices

Introduction

Background

The automated external defibrillator (AED) is a lifesaving device designed to assist individuals experiencing sudden and life-threatening cardiac arrest caused by arrhythmias such as ventricular fibrillation or pulseless ventricular tachycardia. It functions by analyzing the heart rhythm and delivering an electrical shock to reverberate the heart and restore heart rhythm. The AED is made of plastic and has external and internal parts; the external part contains the pad expiration window, latch, status indicator, battery compartment, battery, electrode holder, color display, manual override buttons, shock button, pad or

electrode socket, speaker, and infrared port. The internal parts include the mainboard, display board, speaker, display, shock discharge capacitor, and beeper speaker.

Sudden cardiac arrest is a leading cause of over 350,000 deaths in the United States. The average response time for a first responder is 8 - 12 minutes, and for every minute of delay, the survival rate drops by approximately 10% [1]. The availability of AEDs significantly reduces this response time, underlining its importance in emergency situations.

The AED is a user-friendly instrument that can be operated without the need for specialized training. It is strategically placed in public areas to ensure easy access in the event of an emergency or sudden cardiac arrest. The success of the AEDs is evident in the Chicago Heart Start program, where 22 individuals in cardiac arrhythmia were treated, resulting in 11 successful recoveries. Notably, 6 of these successful treatments were administered by bystanders with no prior AED training.

The AED's circuit board, crucial for its analysis, is constructed from Epoxy FR-4 and measures 254 mm in length, 216 mm in width, and 0.5 mm in thickness. The board is equipped with various components, including a capacitor, microcontroller, flash memory, analog-digital converter, field-programmable gate array (FPGA), processor, audio controller, and indicator. These components, along with others, form the basis of a finite element analysis (FEA) model, which is established based on the board's design specification and components. The governing equation for the experiment is:

$$Mx''+Cx'+Kx=F(t)$$

where M, C, and K represent the mass, damping, and stiffness matrices, respectively.

The goal of the study was to analyze the effect of vibration and thermal experience on the AED, based on its operation. The model underwent various forms of static and dynamic testing of the modeled circuit board. Based on the number and position of the support configurations, the dynamic and vibration properties were analyzed for the modeled circuit board and the rigid board.

This study would assist in obtaining reliable results when the defibrillator is used in mobile ambulances, which experience vibration and road bumping. The AED has been a lifesaving device that is used in conjunction with cardiopulmonary resuscitation (CPR). The results from AED are important as they influence when CPR should be continued or stopped.

Finite element analysis was used for analysis and the model was designed using the Ansys Workbench for the circuit board, which contained integrated circuits. Static and dynamic tests were conducted on the model to determine the bending and damping effects, respectively.

Literature Review

Deformation experienced in electronic components is classified as vibration, shock, and thermal failure. The AED circuit board is a device that experiences failure, which can impact the results obtained during the resuscitation of individuals experiencing cardiac arrest. Generally, most electronic circuit boards experience random vibration rather than ordinary vibration, due to external factors within the vibration environment. Most research on electronics is based on high-cycle fatigue to predict the fatigue life of components experiencing sinusoidal vibration. Fatigue failure under sinusoidal vibration loading has been analyzed by comparing results from vibration failure tests, FEA, and theoretical testing [2].

FEA modeling of the vibration of a printed circuit board (PCB) using rigid boundary conditions has been performed, comparing the modeling results against tests conducted using rigid fixtures to identify the PCB dynamic properties such as natural frequencies [3].

For random vibration fatigue, the circuit board research has extended to soldering by predicting the fatigue life when subjected to random excitation through vibration loading [4]. An experimentally validated vibration fatigue damage model for plastic ball grid array solder joint assemblies was developed by Wu [5] to calculate strain and solder joint survival using a three-band technique.

According to a study by Jadhav et al [6], a virtual simulation method was introduced for onboard charger assemblies using Noise, Vibration, and Harshness analysis. The onboard charger's PCB components' had a first natural frequency of 206 Hz, more significant than the testing's working frequency range. Therefore, resonance effects on PCB components were found to be negligible. Subsequently, frequency response analysis showed that the produced stresses on the onboard charger were minimal. The acceptance criteria confirm the induced von Mises stresses, which fall within acceptable bounds.

As PCB technology becomes more sophisticated, it must balance both electrical and thermomechanical requirements. This study assessed the effects of the constituent layer qualities on the PCB stack-up properties. In the form of cured prepreg, five new FR-4 materials were examined, and thermomechanical parameters, including Young's modulus, glass transition temperature, and coefficient of thermal expansion in the X/Y/Z planes, were measured. The results indicated that as prepreg's glass fiber density rises, the X/Y coefficient of thermal expansion drops, and Young's modulus increases. These values can vary by up to 50%. A reflow temperature of 250 - 260 °C was ideal for lead-free solder; this will minimize the warpage and coefficient of thermal expansion mismatch during processing while increasing reliability. The selection and construction of PCBs with various glass fiber styles, form factors, and dimensions were based on promising FR-4 material options with greater glass transition temperature, reduced coefficient of thermal expansion, and optimized Young's modulus. Prepreg characteristics enhance warpage performance and final PCB properties, regardless of PCB design. While the stack-up design decision mainly influences the PCB's X/Y coefficient of thermal expansion, the PCB's glass transition temperature and Z-coefficient of thermal expansion heavily depend on the resin material's inherent qualities. Both prepreg characteristics and stack-up design decisions impact PCB modulus, as modulus is influenced by copper concentration and distribution. It is recognized that this relationship between prepreg characteristics

and PCB stack-up can improve the understanding of printed circuit board design to achieve higher dependability [7].

In helicopter emergency medical services, a cabin that can swiftly and efficiently transport patients to a hospital is crucial. The quality and safety of the service may be affected by the vibration that patients and crew members experience during transportation, as a medical team uses life-support equipment to maintain the patient's health. However, an incorrect assessment of vibratory level and exposure may result from the airframe's bare dynamical response. The crew, patients, and medical supplies, through interfaces like seats, handles, stretchers, and flexible supports, dynamically interact with the helicopter. Therefore, to create a low-vibration helicopter emergency and medical service vehicle, the coupled helicopter-interface-subject system must be the subject of thorough numerical analysis. It should be possible to run the analysis effectively and efficiently across many possible configurations to achieve optimal positioning. Formulating high-fidelity rotorcraft aeroservoelasticity, connecting additional dynamical systems that represent the dynamics of humans and equipment, and calculating the vibration performance of the resulting models should all be possible with a viable tool. An efficient method for assessing the vibratory performance of medical helicopters was presented. The strategy is demonstrated on a medium-sized helicopter by adding dynamical models of a human resting on a seat, a recumbent person lying on a stretcher, and medical equipment mounted on flexible supports at its ends [8].

The study by Oon et al [9] investigated the warpage of both, a minor PCB with only one side and a large PCB with multiple layers using Shadow Moiré measurement and FEA. Primarily due to the absence of an initial warpage, simulation results from both printed circuit boards are lower than experimental results. At temperatures below the sample's glass transition temperature, the disparity is reduced to 24% for the single-sided, small PCB and 15% for the multilayer, large PCB, respectively. The PCB conduct could not be reflected in the reproduction at the above glass transition temperature. Their findings demonstrate that the simplified model based on the copper content of each layer can be used to estimate the sample's warpage for the multilayer, large PCB. When the copper content of each layer in the sample is increased to 100%, it is estimated using the same method that the PCB warpage would increase by 25%. In contrast, halving the copper content of each copper layer results in a 21% decrease in warpage.

The study by Yun et al [10] is a comprehensive examination of how well AEDs function during transportation in a moving ambulance. The researchers aimed to determine whether the motion of an ambulance affected the AED's ability to detect and treat cardiac arrhythmias accurately. The study found that the performance of AEDs was reliable even while in motion, though certain variables, such as speed and road conditions, may affect their efficiency. This thorough research instills confidence in the reliable use of AEDs in emergency medical services during transport.

The study by Wang et al [11] explored the challenges of managing vibration levels in life-support systems used in medical helicopters. Their study focused on how excessive vibrations during air transport can affect patient care and the functionality of medical devices, such as ventilators and defibrillators. Using advanced simulation techniques, the researchers modeled how vibrations from the helicopter's frame are transmitted to onboard equipment and patient positioning systems. They identified optimal configurations to reduce mechanical stress on life-support systems, ensuring improved patient stability and care during transport. The findings highlight the importance of minimizing vibrations to enhance the safety of medical air evacuations.

Recent advancements in random vibration fatigue research, especially in AEDs, have seen significant progress with enhanced FEA methods [12], including improved fatigue life predictions for solder joints within PCBs by simulating real-world vibrational conditions like those experienced in ambulances. Their study introduced advanced algorithms to represent random excitation forces better, by improving fatigue life predictions' reliability. By refining input variables, their study enabled more accurate testing and earlier detection of potential failure points, thus enhancing the durability and safety of critical medical devices such as AEDs in unpredictable environments.

The studies by Caffrey et al, Chen et al, and Jespersen et al [13-15] investigated the functionality and reliability of AEDs in critical life-saving situations but from different perspectives [15]. Focusing on the operational status of AEDs registered for public use revealed that many AEDs were nonfunctional when needed for out-of-hospital cardiac arrests (OHCA), due to issues such as dead batteries or missing components, underscoring the importance of regular maintenance. Similarly, Chen et al [14] further evaluated the performance of AEDs under extreme environmental conditions, such as high altitudes and fluctuating temperatures, and found that these conditions lead to decreased battery life and impaired shock delivery efficiency. This study highlights the need for innovation in AED technology, mainly to ensure functionality in challenging environments, such as mountainous or cold-weather regions. Together, these studies emphasize the critical need for regular AED maintenance and design improvements to ensure reliability in varied emergency conditions, which is vital for improving survival rates in OHCAs.

Olalere and Choi [16] explored innovative voltage regulation techniques to improve the performance of electrosurgical devices, particularly of the hot snare polypectomy tool. Their work aligns with this study's investigation into circuit board deformation in medical devices. The voltage control techniques developed by them helped optimize energy output during surgical procedures, thereby reducing thermal damage to surrounding tissues and improving patient outcomes. This study provides significant context for understanding how optimized electronic control systems can enhance the durability and functionality of medical devices, especially under stress or thermal conditions, as seen in the AED analysis. This research is directly relevant to ongoing work on improving device reliability in adverse conditions, such as high-frequency vibration environments and extreme thermal variations. Their focus on energy modulation parallels the need for efficient heat

dissipation and improved circuit board resilience, which are key concerns in designing and optimizing critical medical equipment like AEDs.

Studies by Jonsson et al and Sarkisian et al [17,18] emphasize the critical role of AEDs in out-of-hospital cardiac arrest situations but from complementary angles. They highlighted the significant survival benefits of integrating AEDs into first responder systems, especially in public spaces. Their findings reveal that timely access to AEDs, typically within the first few minutes of cardiac arrest, drastically improves survival outcomes. Integrating technology, such as real-time location services, enhances the ability of first responders to locate and use AEDs more effectively, leading to faster defibrillation times and better resuscitation outcomes. This study underscores the importance of placing AEDs in high-traffic areas to maximize accessibility and efficacy [18]. Further examination of the effectiveness of AED deployment in different public and private settings revealed disparities in AED availability, with urban public areas being better equipped than residential locations, where most OHCA incidents occur. They emphasize the necessity for strategic AED placement and increased accessibility, particularly in residential and low-traffic areas, to close the gap in survival rates. Both studies converge on the idea that improved AED access and integration with first responder systems are essential for enhancing OHCA outcomes, primarily through strategic placement and the use of technology.

Several studies have explored the integration of drones in delivering AEDs to OHCA locations [19-21]. They focused on the experiences of dispatcher nurses managing these drone deployments, highlighting benefits such as faster AED delivery in rural areas, while also noting challenges such as synchronizing drone arrival with bystander readiness [20]. The authors emphasize regulatory, technical, and logistical barriers in coordinating drone technology with emergency medical services (EMS), identifying the need for standardized protocols [20]. Their study also notices that countries such as Sweden and Canada have made more significant progress in integrating drones into emergency systems than the United States. Schierbeck et al [21] have evaluated the effectiveness of drone-delivered AEDs in Sweden and showed that drones consistently arrive faster than ambulances in real-life OHCA scenarios. However, they also highlighted operational limitations such as weather conditions. These studies underscore the potential effect on the AED's performance and efficiency in improving emergency response, particularly in rural or hard-to-reach areas. However, they also call for greater coordination, infrastructure improvements, and policy development to maximize their impact.

Several studies have collectively examined different strategies to improve defibrillation outcomes in OHCA locations [22-24]. They compared the effectiveness of on-site bystanders performing defibrillation with dispatched volunteer responders, highlighting that while on-site bystanders provide more immediate aid, dispatched responders often have better training and equipment. They suggest a hybrid approach where bystanders initiate CPR and responders handle defibrillation, increasing survival chances [23]. They also focus on the challenges of increasing defibrillation rates in home-based OHCAs, noting the lack of access to AEDs and limited public awareness as key barriers. Their recommendations include expanding AED distribution in residential areas and increasing public training [24]. An analysis of the effectiveness of volunteer responder systems in rural and urban areas found that volunteers significantly improved response times and survival rates in less densely populated areas, filling critical gaps where traditional emergency services may face delays. These studies underscore the need for integrated strategies, combining public training, AED distribution, and volunteer responders to improve survival rates across different geographic settings.

Methods

Model Preparation

The modeled circuit board components used in this study include capacitors, microcontrollers, flash memory, analog-digital converters, FPGAs, processors, audio controllers, batteries, and transistors.

The AED is a model comprising different components for analysis. The baseboard measures 254 mm × 216 mm.

The capacitors are cylindrical, with a length of 40 mm and a diameter of 35 mm. The microcontroller is 10 mm × 10 mm × 1.4 mm. The battery design is 34.5 mm in length and 17 mm in diameter.

The materials used in the components along with their respective Young's modulus and Poisson ratio are presented in Table 1.

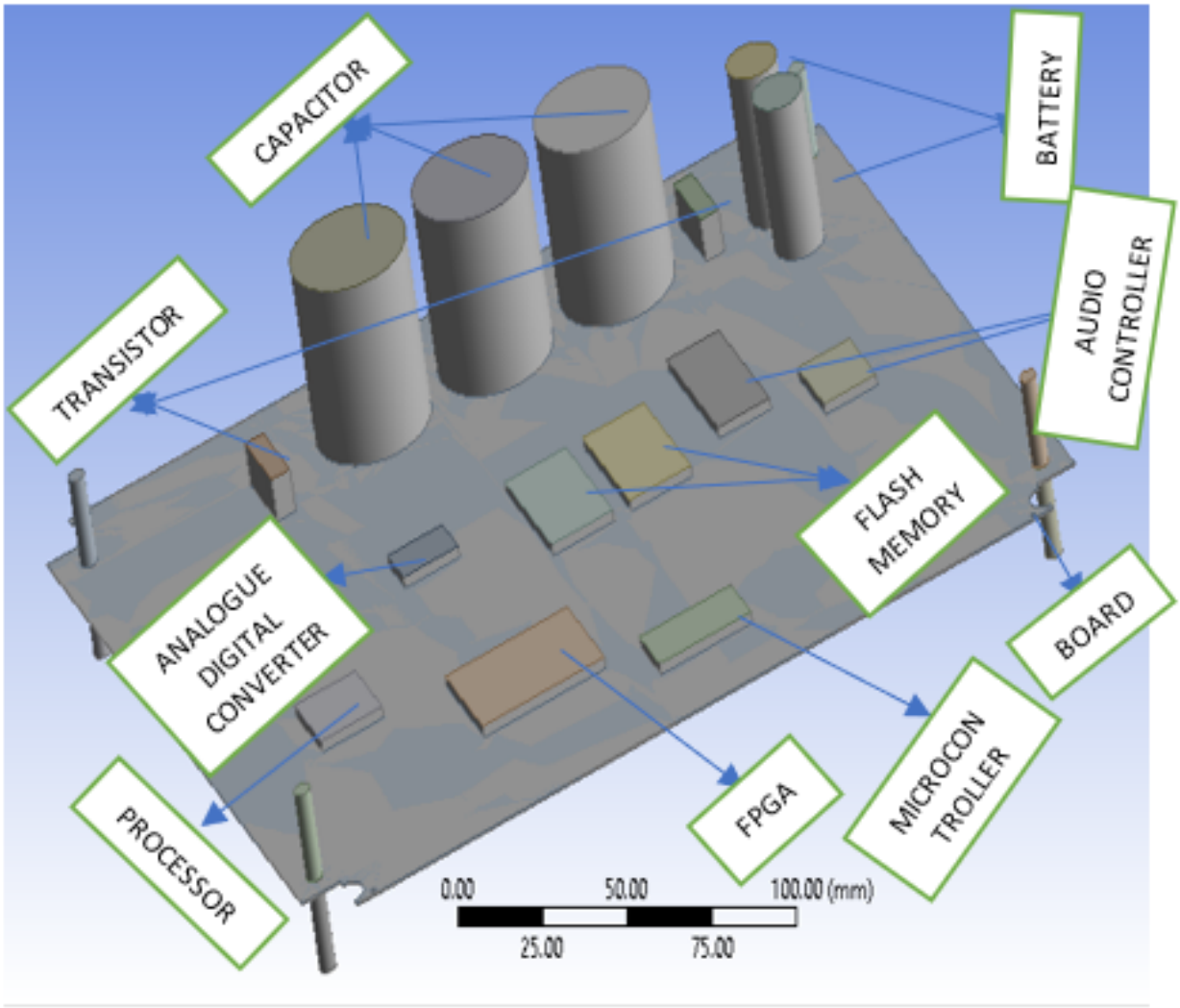
Table . Numerical constant of component.

Component	Material	Young's modulus (GPa)	Poisson ratio	Thermal conductivity (W/mK)
Board	FR4 epoxy	24	0.118	0.81
Capacitor	Tantalum	175	0.34	54.4
Microcontroller	Copper	117	0.34	385
Flash memory	Polystyrene	3250	0.34	0.033
Analog digital converter	Silicon	140	0.275	150
FPGA[a]	Silicon	140	0.275	150
Processor	Silicon	140	0.275	150
Audio controller	Copper	117	0.34	385
Battery	Lithium	3.17E-05	0.355	5.4
Transistor	Silicon	140	0.275	150

[a]FPGA: field-programmable gate array.

The modeled AED, developed using Ansys 2020 through Workbench, is a testament to our thorough approach to design. The model circuit board, designed based on the specific dimensions of the board and its components, further underscores our commitment to precision. Four fixed supports for the AED were fixed to the plastic casing of the AED, ensuring stability and reliability. The FEA, a powerful tool for deformation analysis and thermal effect assessment, was used to conduct a comprehensive analysis of the circuit board and its components, providing reassurance about the robustness of our design. The FEA model, a visual representation of our thorough analysis, is presented in Figure 1. The boundary condition was set at the four edges of the modeled circuit board, which are fixed as rigid bodies, mirroring a typical AED.

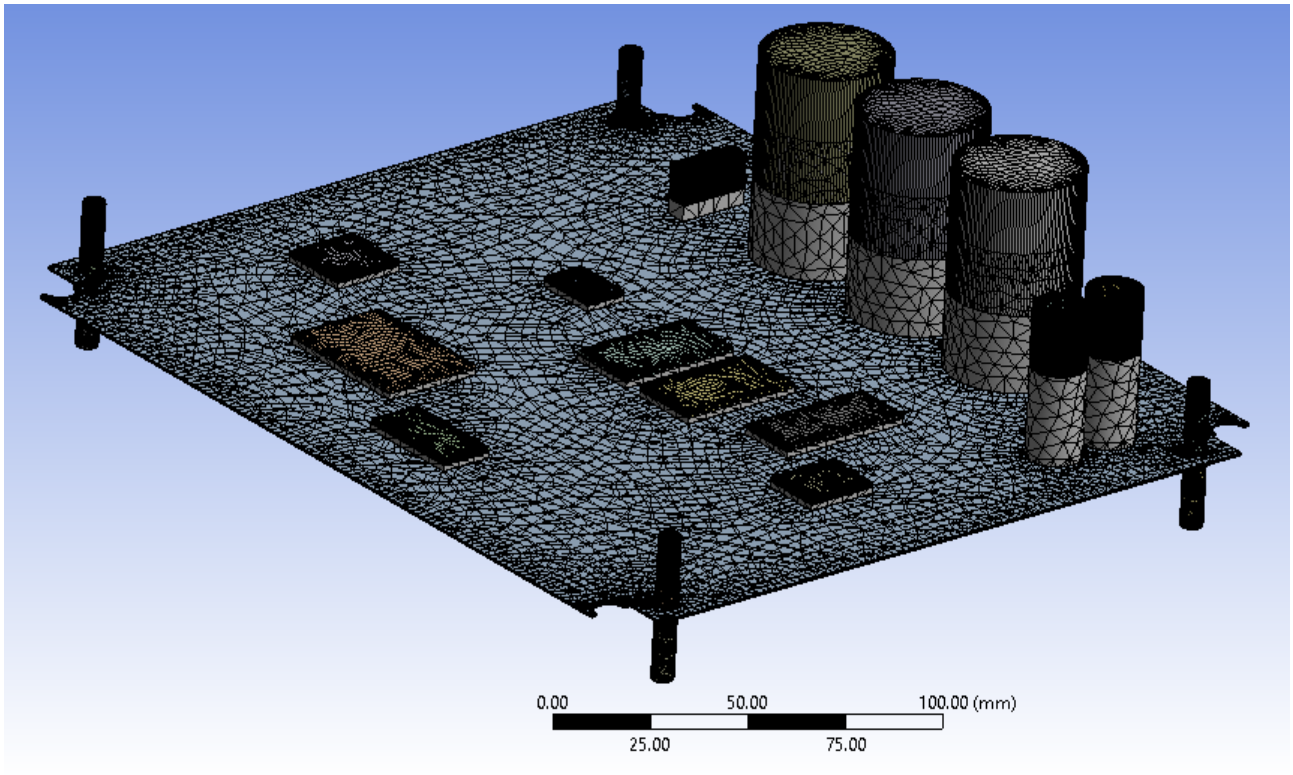
Figure 1. Model of the circuit board. FPGA: field-programmable gate array.



Mesh Selection

The modeled circuit board was meshed to verify the stress discontinuity of the member components attached to the board. The mesh model produces 90,371 nodes and 59,671 elements from program-controlled order (Figure 2).

Figure 2. The mesh of the circuit board.



Ethical Considerations

This study did not require ethics board approval, as it did not involve human participants, identifiable personal data, or biological materials. No intervention or interaction with individuals occurred, and all data analyzed were obtained from simulation analysis. Therefore, an application to an institutional review board was not necessary, in accordance with policies and national regulations.

Results

Deformation Analysis for 4-Member Support

The circuit board undergoes different deformation at different parts of the board. The maximum deformation is in the middle of the board, at a peak of 33.141 mm. Likewise, the entire board experiences edge bending, as shown in Figure 3. Taller components deform faster, causing damage to the circuit board.

The rotational deformation, measured with precision, is at its largest at the z-axis, with transverse at a 10% rate before converging at 60%. The x-axis shows a precise deformation at 10% - 60% of the time series. The y-axis deformation, also measured with precision, ranges between 10% - 80% of the time series before finally converging, as shown in Figure 3.

According to Table 2 and Figure 4, approximately more than 30% of the effective mass contributed to the mode in the X direction. In comparison, 50% and 55% of effective masses contributed to the Y and Z directions, respectively.

Figure 3. Static structural deformation for 4-member support.

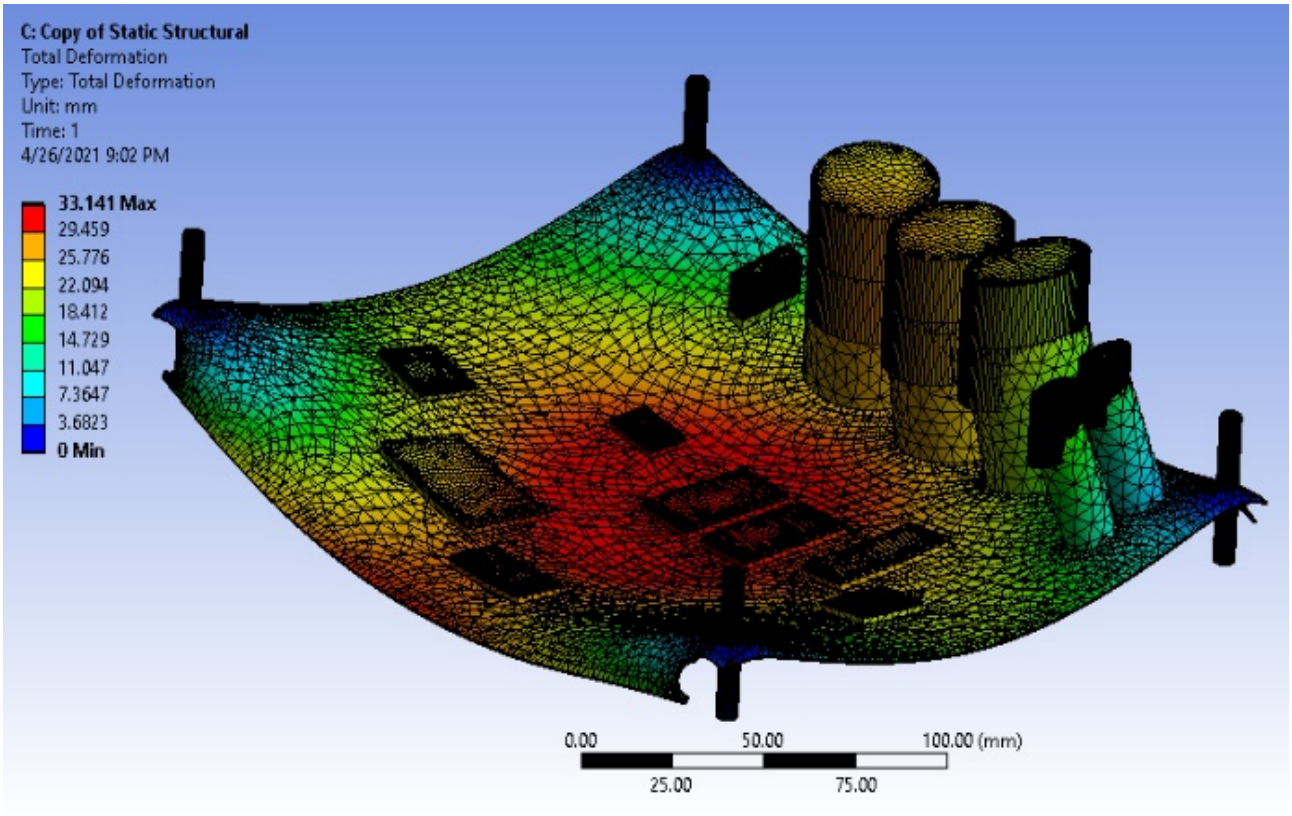


Figure 4. Effective mass to frequency for 4-member support.

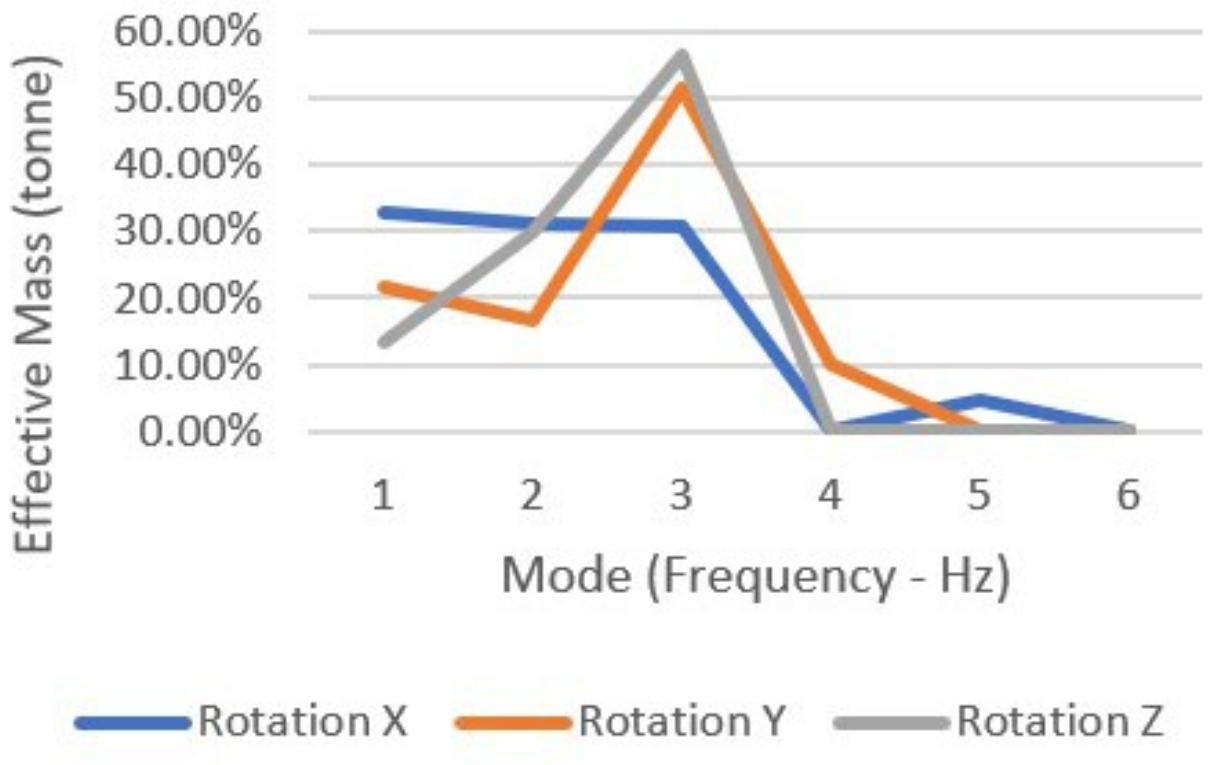


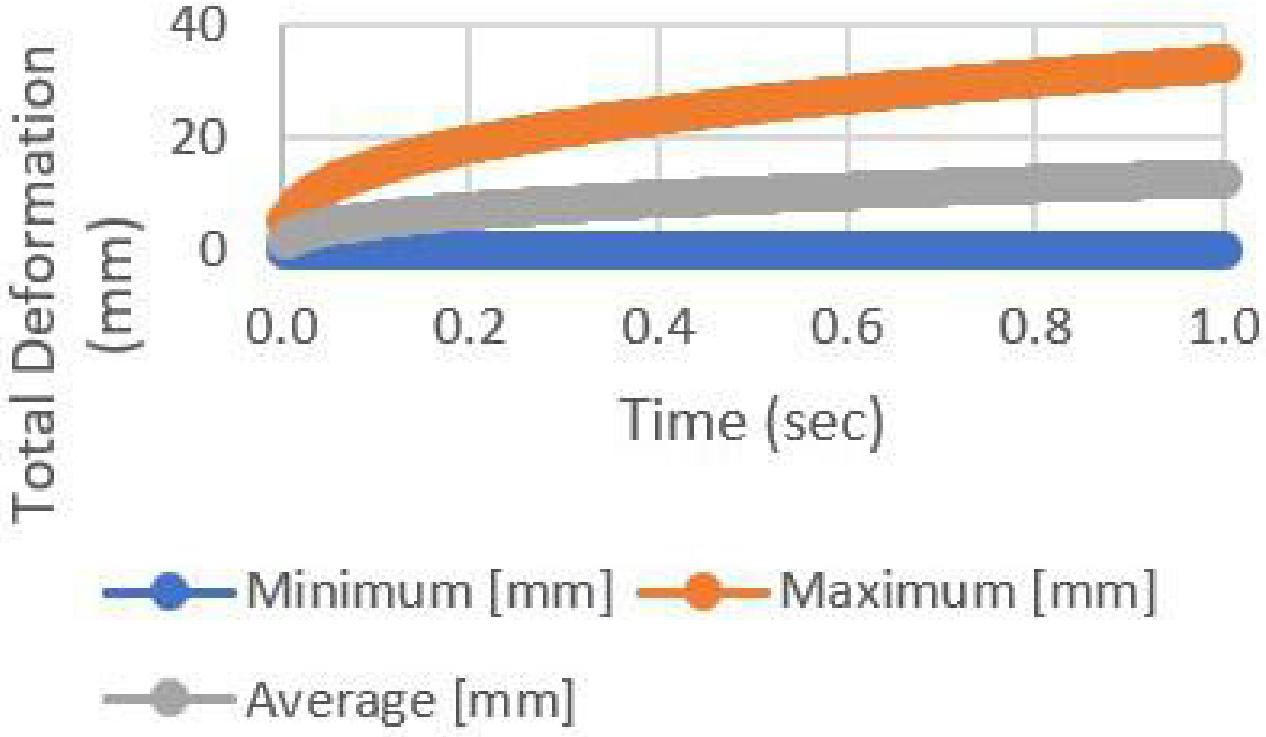
Table . Mass participation for 4-member support.

Mode	Rotation X	Rotation Y	Rotation Z
1	33.669	15.71	21.22
2	3.20E+01	1.21E+01	4.72E+01
3	3.18E+01	3.71E+01	8.91E+01
4	1.61E-01	7.13E+00	1.33E-01
5	4.98E+00	9.73E-02	7.10E-03
6	2.18E-03	3.70E-03	4.09E-03
Overall	102.6497	72.17238	157.6688

Our research on the structural fatigue of the circuit board has yielded clear and important findings. We found that the center of the board, with less support than the four edges, experiences more visible deformation. This understanding is crucial for the development of more robust electronic components.

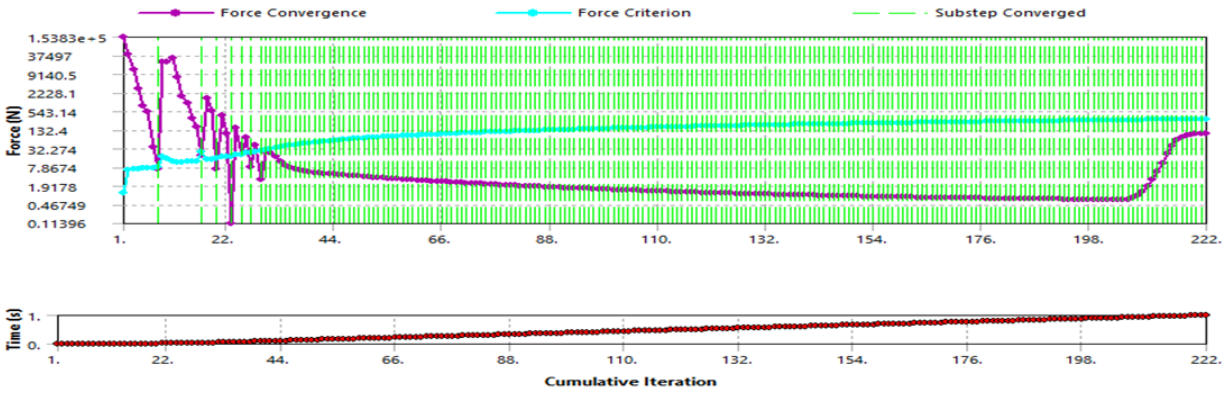
Figure 5 shows the rate of deformation at an increasing level, which shows the time-to-failure for the specific components, such as resistors and capacitors, during the vibration test.

Figure 5. The total deformation rate of the circuit board.



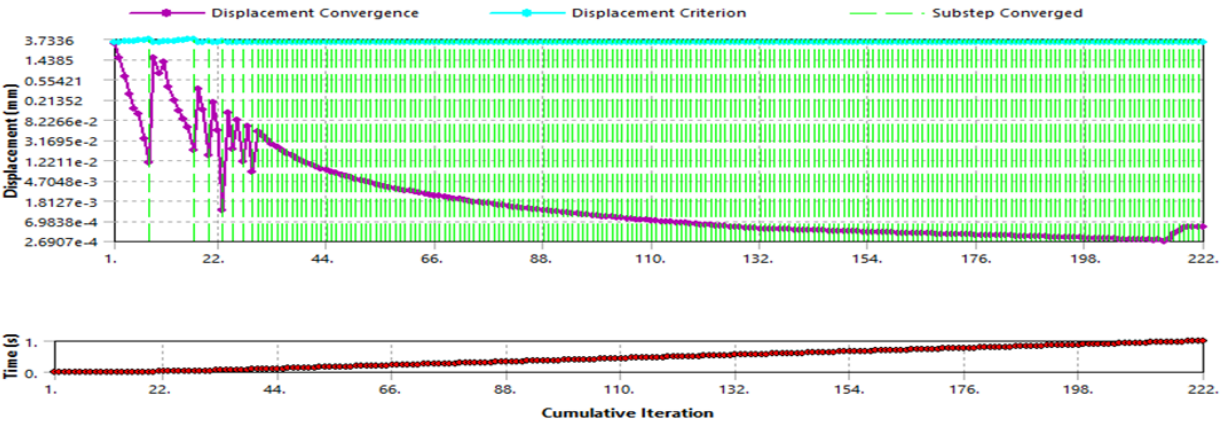
For the force convergence in Figure 6, the substep converged towards the iteration end point. The convergence experienced longer iterations, allowing for the load to be evenly distributed, as seen by the substep. At 20% of the loading, normal stiffness was lowered to improve the analysis results.

Figure 6. Force convergence.



From the displacement convergence in Figure 7, the standard stiffness was maintained towards the analysis's tail end, indicating a consistent and accurate deformation rate from the 15% iteration.

Figure 7. Displacement divergence.



The moment of convergence in Figure 8, with the mesh refinement leading to node increment, demonstrates a uniformly converging analysis, ensuring the reliability of the results.

Figure 8. Moment convergence.

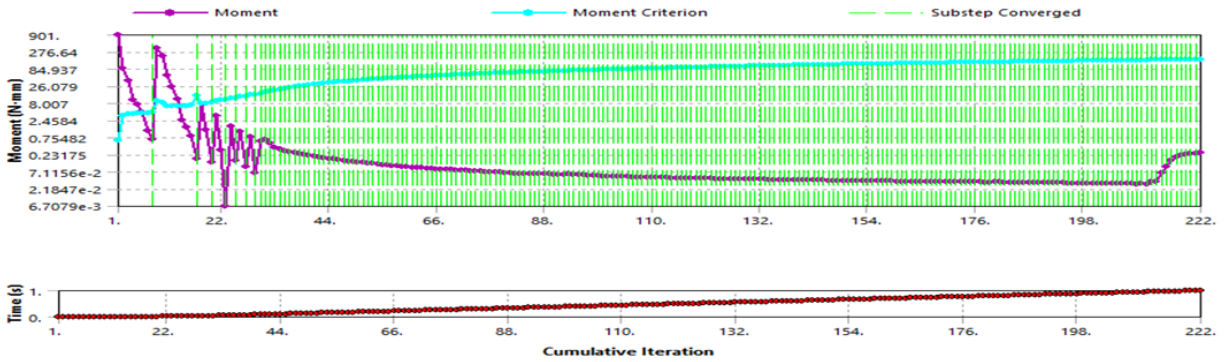


Figure 9 shows the deformation for an unprestressed condition. The deformation, which was more noticeable at the location of the capacitors, led to a significant distortion in the circuit board's shape. This distortion could potentially affect the performance of the circuit board, highlighting the importance of considering deformation in the design process.

The modal analysis used to investigate the vibration on the circuit board is used to evaluate the natural frequencies, as shown in Figure 10.

Figure 9. Unprestresses modal total deformation.

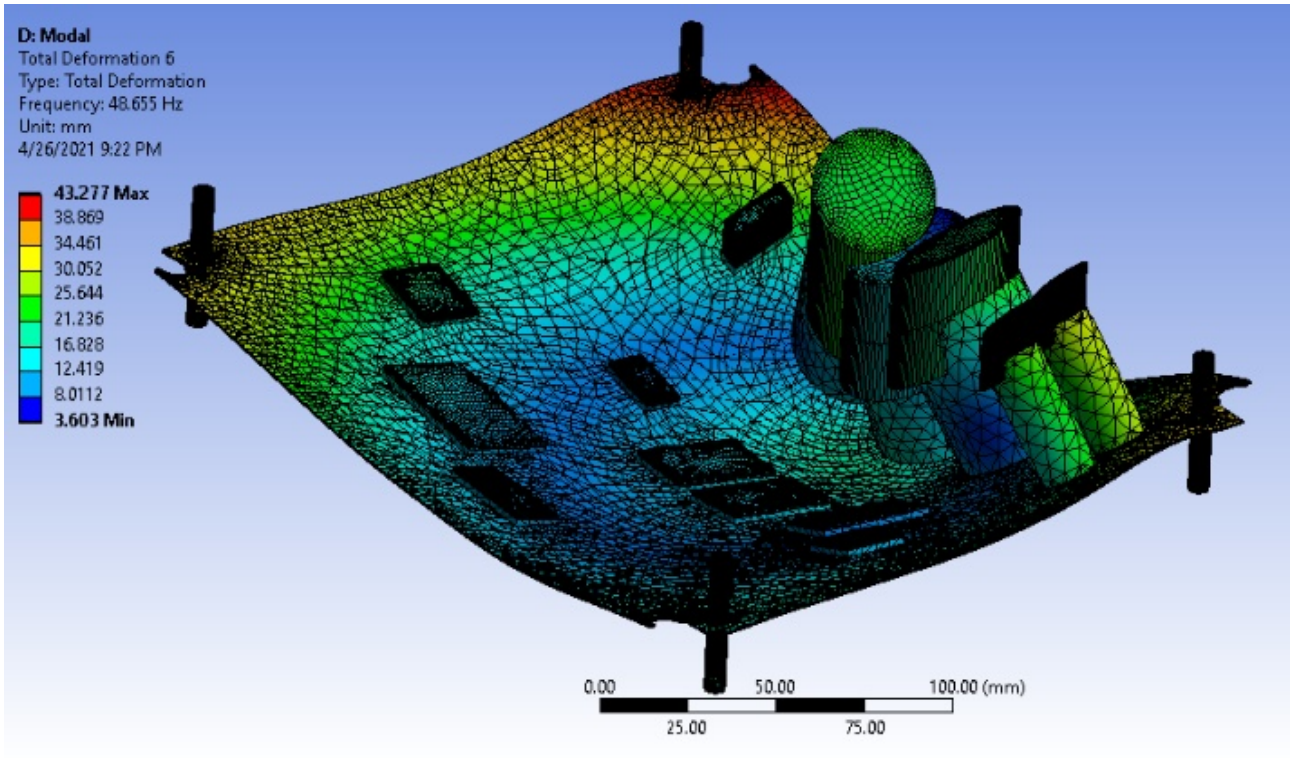
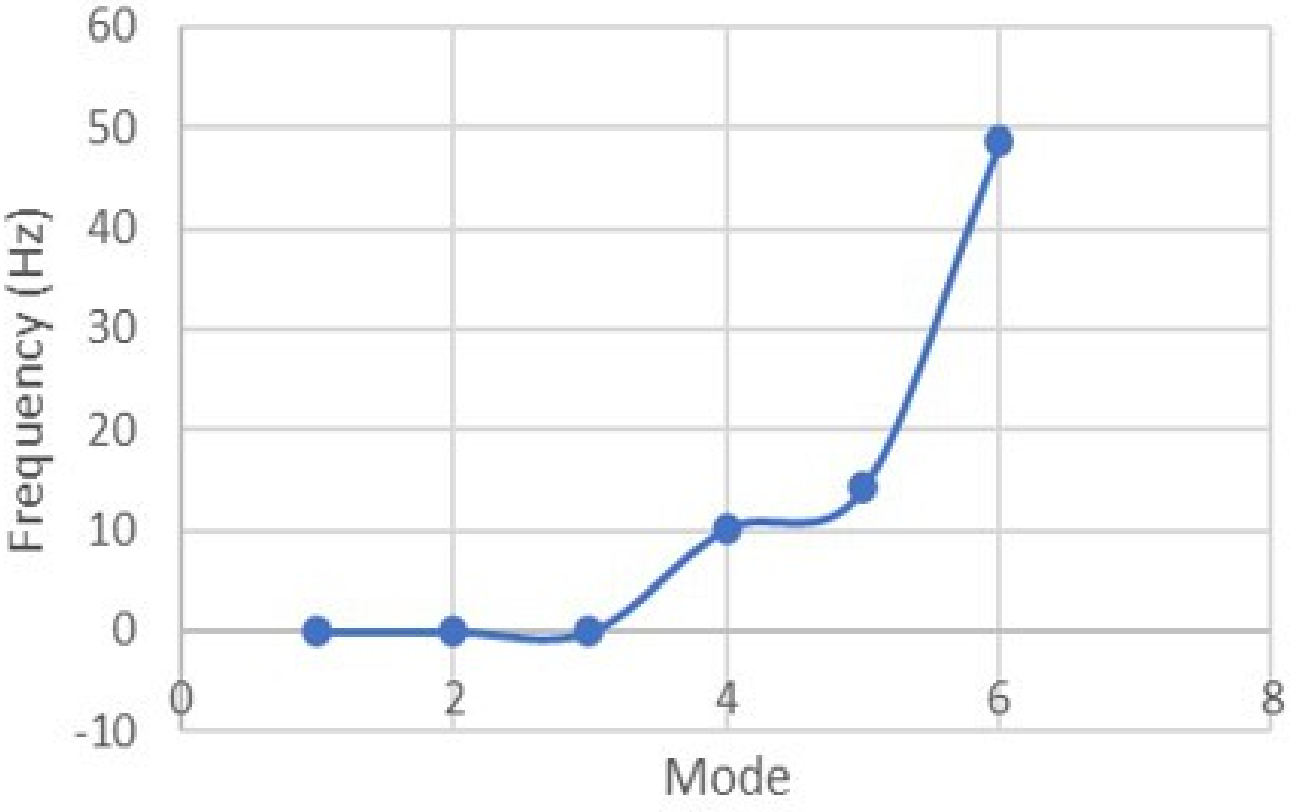
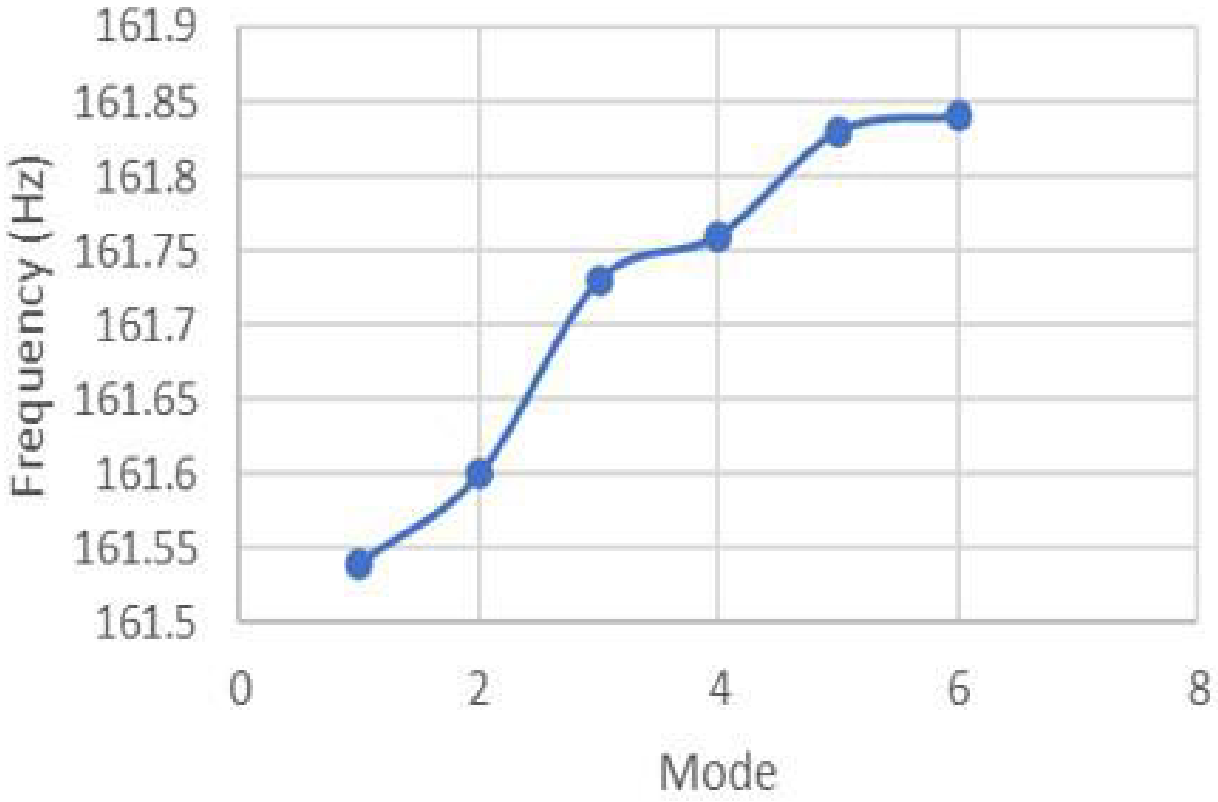


Figure 10. Unprestressed modal analysis.



The prestressed analysis assists in improving the results obtained in the unprestressed process by modifying the stiffness to reduce the natural frequency inadequacies. This provides better and improved results for the simulation as the analysis was refined to give better frequencies as against the unprestressed, as seen in Figure 11.

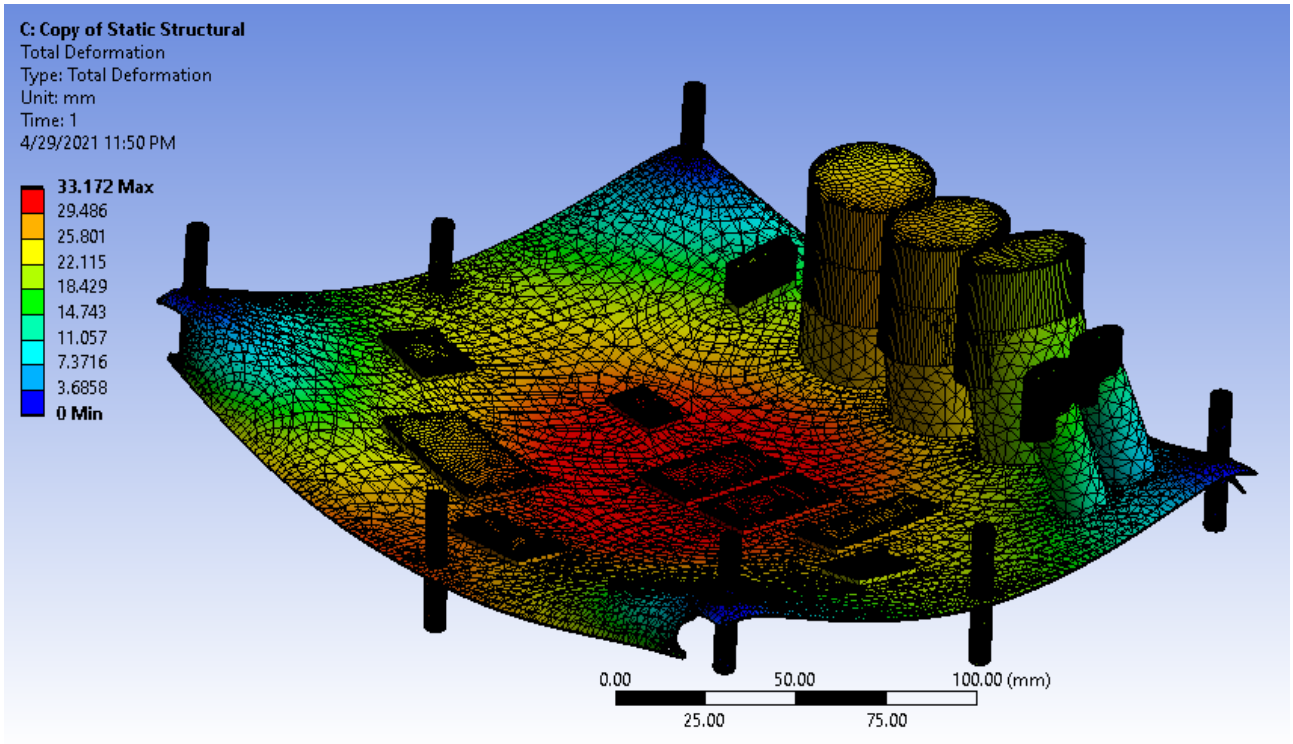
Figure 11. Prestressed modal analysis.



Deformation Analysis for 8-Member Support

The modeled circuit board was reinforced with more support members to improve its deformation effect. The deformation peaked at 33.172 mm with minimum deformation at the board's edge, as shown in Figure 12.

Figure 12. Static structural deformation for 8-member support.



As seen in Figure 13 and Table 3, more than 95% of the effective mass was the participating mode in the Z direction, 60% in the X direction, and slightly below 60% of the mode's effective masses in the Y direction.

Figure 13. Effective mass to frequency for 8-member support.

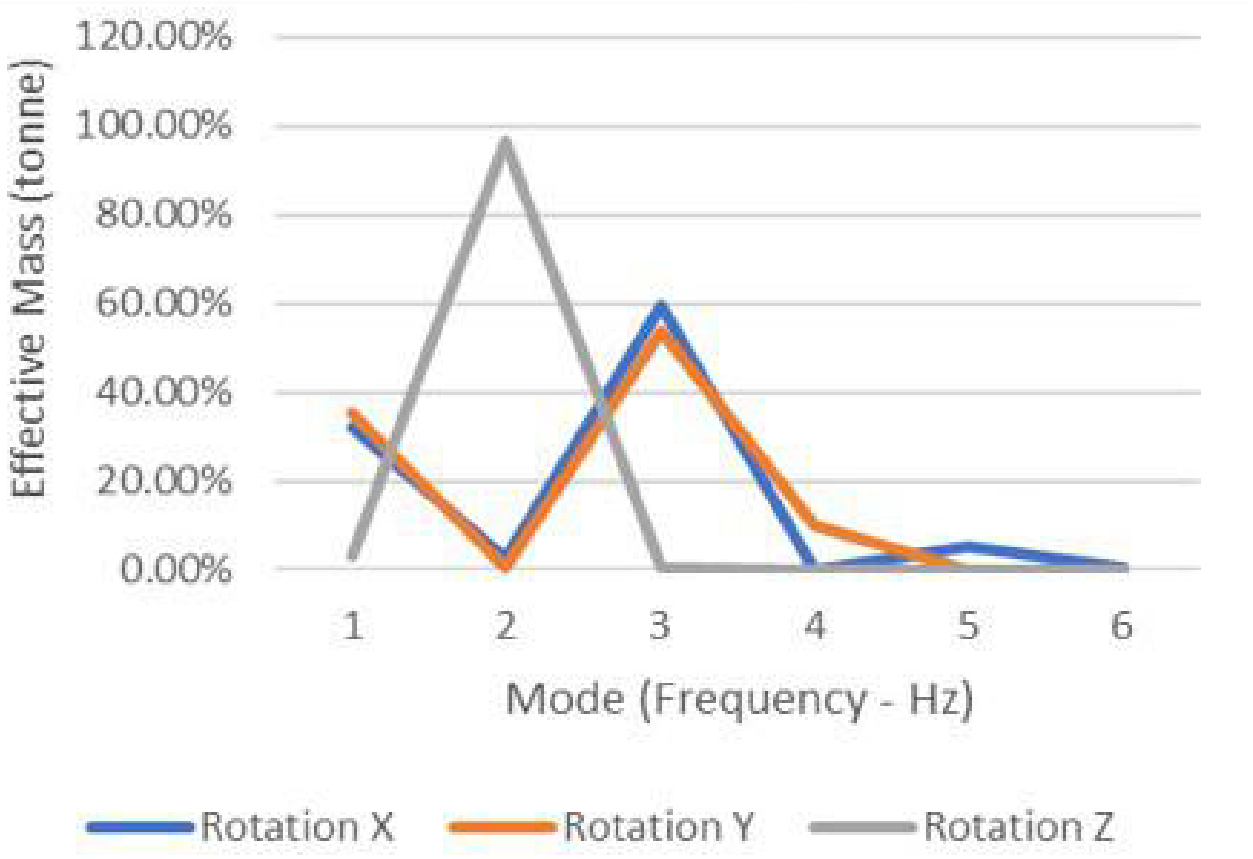
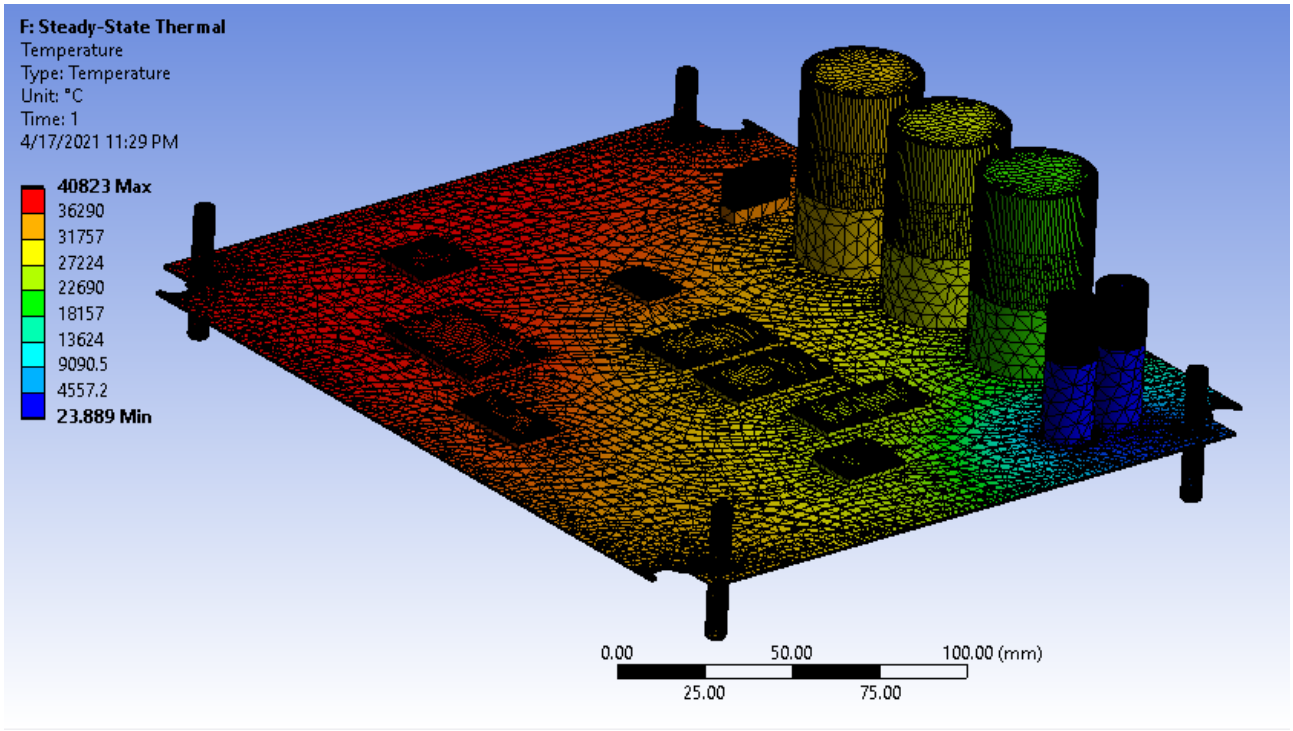


Table . Mass participation for 8-member support.

Mode	Rotation X	Rotation Y	Rotation Z
1	33.442	25.723	5.2324
2	2.67E+00	4.11E-01	1.53E+02
3	6.17E+01	3.95E+01	4.72E-01
4	6.09E-03	7.47E+00	2.01E-01
5	5.24E+00	1.06E-02	1.01E-03
6	3.00E-01	3.94E-03	8.40E-03
	103.3538	73.10313	158.784

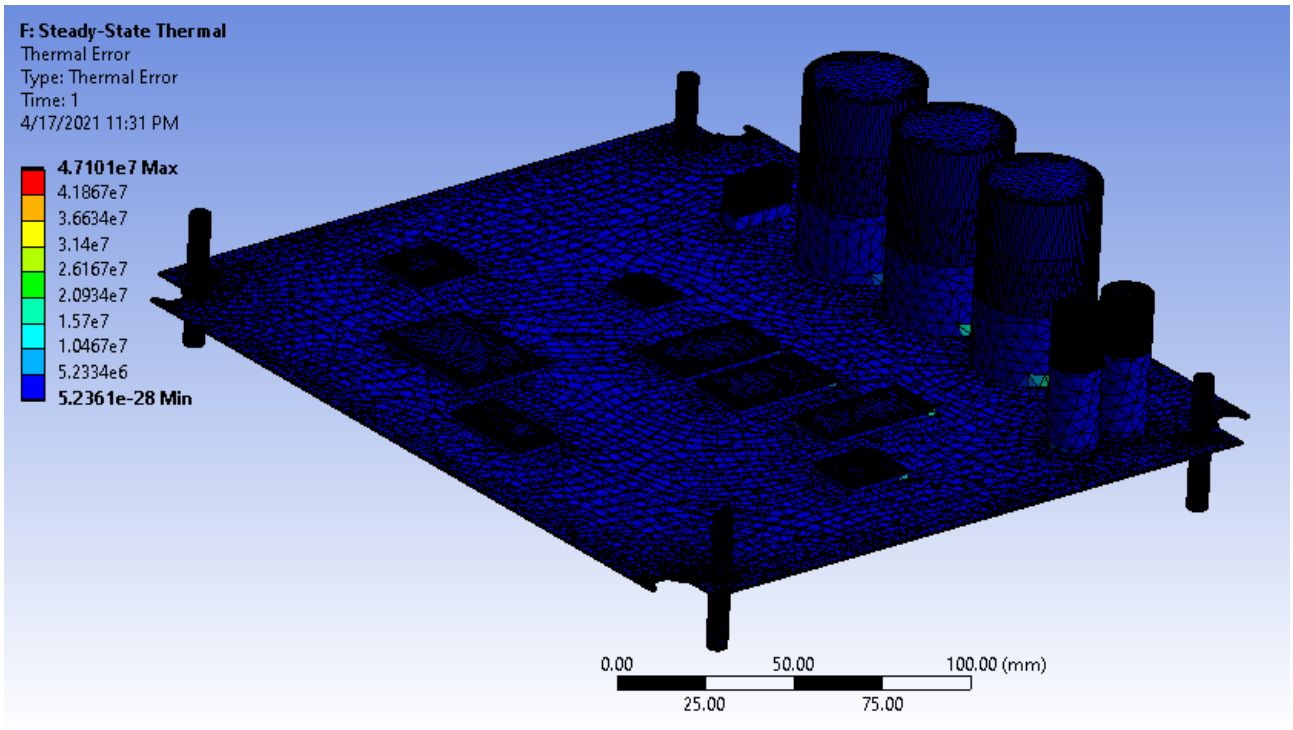
The blue part of the circuit board shown in Figure 14 is the battery of the modeled circuit board, which is used to power the board. The temperature is distributed at this point, a critical factor that significantly impacts the performance of the entire circuit board. The heat dissipation to the circuit was high, more than 40,000 °C compared to the initial temperature of 23 °C based on the surface area and dissipation rate.

Figure 14. Thermal state of the circuit board.



The battery's thermal error, as shown in Figure 15, is higher than expected. This is due to the temperature dissipated within the components encountered by the circuit board. The temperature effect gives a higher thermal error at a short time interval, which will lead to a rapid and concerning rate of the circuit board's deformation.

Figure 15. Thermal error of the circuit board.



Discussion

Principal Findings

This study made a unique contribution to the field of medical device design by employing FEA analysis to evaluate the modal and thermal responses of an AED. The use of Ansys Workbench for static structural, modal, and steady-state thermal analysis provided novel insights into the relationship between modal frequencies and thermal effects on the AED's circuit board, focusing on component deformation and heat dissipation.

The FEA model demonstrated natural frequencies consistent with similar experimental models. This alignment indicates the

robustness of the model but also highlights the influence of boundary conditions and the smeared property approach in the simulation. The boundary conditions may have contributed to the observed differences in frequency ranges between the FEA and experimental models, reflecting the inherent complexity of capturing real-world constraints within a simulation environment.

Overall, the findings underscore the crucial role of considering thermal and mechanical factors in the design of medical devices like AEDs, particularly for ensuring reliability under varying operational conditions. The implications of this study could significantly influence the field of medical device design, leading to the development of more robust and reliable AEDs. The study's findings could shape the design and development of AEDs, resulting in devices that are better equipped to withstand the challenges of real-world use, thereby improving patient outcomes and enlightening the field.

Strengths

First, including both prestressed and unprestressed modal analysis adds a layer of accuracy to the model, which improves the simpler FEA models used in previous research. The 0.0003% error difference in the prestressed analysis results is significant as it ensures the model's accuracy, which is crucial for predicting AED performance. This level of precision is a substantial improvement over previous models, indicating the potential for more reliable AED designs.

The study's ability to isolate the battery's thermal effects and provide practical design recommendations, such as using separate boards or cooling mechanisms, equips engineers and researchers with actionable solutions for improving the durability of AEDs. These recommendations, such as using separate boards for the battery to reduce thermal stress on the main circuit board or implementing a cooling mechanism like a fan to manage the heat effectively, can be directly implemented in device designs, thereby enhancing the reliability of AEDs.

Lastly, the research leverages detailed meshing and refined convergence methods, ensuring uniformity in the results and yielding more reliable data for future applications in device design. This emphasis on methodological rigor should instill confidence in the audience about the reliability of the study's data and its potential for future applications in device design.

Comparison to Prior Work

The results are consistent with prior studies in the field, such as [10,11], which similarly demonstrated the adverse effects of vibrations and temperature fluctuations on medical devices like AEDs. However, this study extends previous findings by incorporating a prestressed analysis that improved natural frequency estimations, showing a minimal percentage error of 0.0003%. This precision in the FEA model instills confidence in the audience, representing an advancement in predicting deformation and fatigue, especially compared to earlier studies that primarily focused on random vibration fatigue or single-mode vibration analysis.

Limitations

During this study, the detailed joints between the components and the circuit board are not considered; the emphasis is on the face-to-face contact of the components with the circuit board.

Second, the parametric iteration method used to determine the damping varies from 0.001 to 0.005 in a step of 0.005 sec. This limitation could lead to discrepancies between the model's predictions and real-world outcomes, especially under varying temperatures and vibrational stresses that are not uniformly distributed in practice.

Third, while the study used prestressed analysis and provided accurate frequency improvements, the material properties of the components, such as capacitors and batteries, were modeled with assumptions that may not fully capture the complexity of real-world materials. For instance, variations in thermal expansion coefficients and conductivity might behave differently under prolonged or extreme use, leading to more pronounced deformations than simulated.

Lastly, the study did not extend its analysis to long-term fatigue testing of components under cyclic loading. Without fatigue life predictions, it is unclear how the PCB and its components would withstand continuous use over time, especially in emergency medical environments where reliability is critical.

Conclusion

The conclusion of this study highlights the importance of prestressed analysis in improving the accuracy of vibration analysis for PCBs used in critical medical devices like AEDs. The presence of zero frequencies, which are attributed to the rigid body modes, introduces superfluous effects that were effectively minimized through weak springs, improving the overall accuracy of the vibration response. This step enhanced the natural frequency results, as shown by the prestressed analysis, which produced a negligible error of 0.0003%, demonstrating its precision.

The stiffness of the modeled circuit board was unevenly distributed due to the varying materials and components mounted on it, such as capacitors, which exhibited pronounced deformation. It is recommended to use flat capacitors of lower height to mitigate this issue, as they are less prone to deformation under vibration stress. Additionally, the significant heat dissipation from the lithium battery, which has a high specific heat capacity, was identified as a potential source of thermal stress, affecting the PCB's long-term reliability. The study suggests using a dedicated cooling system, such as a fan, or placing the battery on a separate board to manage the heat effectively.

The findings suggest that the prestressed analysis method, enhanced thermal management strategies, and design modifications may improve the durability and performance of PCBs in real-world operating conditions. This could be important for the reliability of medical devices in environments like ambulances, where continuous vibration and thermal fluctuations occur.

Acknowledgments

The author wishes to thank Professor Rahman Mosfequr.

Data Availability

All data generated or analyzed during this study are included in this published article.

Authors' Contributions

Conceptualization: SOO
Formal analysis: SOO
Methodology: SOO
Validation: SOO
Writing – review & editing: SOO

Conflicts of Interest

None declared.

References

1. AED - using an AED - what is an AED? American Red Cross Training Services. URL: https://www.redcross.org/take-a-class/aed/using-an-aed/what-is-aed [accessed 2021-04-10]
2. Chen YS, Wang CS, Yang YJ. Combining vibration test with finite element analysis for the fatigue life estimation of PBGA components. Microelectron Reliab 2008 Apr;48(4):638-644. [doi: 10.1016/j.microrel.2007.11.006]
3. Wu J, Zhang RR, Radons S, Long X, Stevens KK. Vibration analysis of medical devices with a calibrated FEA model. Comput Struct 2002 May;80(12):1081-1086. [doi: 10.1016/S0045-7949(02)00067-6]
4. Alyafawi A, Yu D, Park S, et al. Reliability assessment of electronic components under random vibration loading. Presented at: Electronic Components and Technology Conference; May 26-29, 2009; San Diego, CA. [doi: 10.1109/ECTC.2009.5074049]
5. Wu ML. Vibration-induced fatigue life estimation of ball grid array packaging. J Micromech Microeng 2009 Jun 1;19(6):065005. [doi: 10.1088/0960-1317/19/6/065005]
6. Jadhav S, Kalurkar S, T S S. Vibration analysis of an on-board charger assembly for electrical mobility. ARAI J Mobi Tech 2023;3(3):658-665. [doi: 10.37285/ajmt.3.3.3]
7. Liu T, Devarajan M. Influence of prepreg material properties on printed circuit board (PCB) stack-up. Presented at: 2022 IEEE 72nd Electronic Components and Technology Conference (ECTC); May 31 to Jun 3, 2022; San Diego, CA. [doi: 10.1109/ECTC51906.2022.00354]
8. Tamer A, Muscarello V, Masarati P, Quaranta G, Politecnico Di Milano. Vibration rating of medical helicopters. Presented at: Vertical Flight Society 74th Annual Forum & Technology Display; 2018; Phoenix, AZ. [doi: 10.4050/F-0074-2018-12755]
9. Oon SJ, Tan KS, Tou TY, Yap SS. Warpage studies of printed circuit boards with shadow moiré and simulations. Presented at: IEEE 38th International Electronics Manufacturing Technology Conference (IEMT); Sep 4-6, 2018; Melaka, Malaysia. [doi: 10.1109/IEMT.2018.8511790]
10. Yun JG, Jeung KW, Lee BK, et al. Performance of an automated external defibrillator in a moving ambulance vehicle. Resuscitation 2010 Apr;81(4):457-462. [doi: 10.1016/j.resuscitation.2009.12.031] [Medline: 20122777]
11. Wang Y, Fang Y, Li L, Zhang D, Liao WH, Fang J. Dynamic modeling and vibration suppression of a rotating flexible beam with segmented active constrained layer damping treatment. Aerospace 2023;10(12):1010. [doi: 10.3390/aerospace10121010]
12. Salhi RA, Fouche S, Mendel P, et al. Enhancing Prehospital Outcomes for Cardiac Arrest (EPOC) study: sequential mixed-methods study protocol in Michigan, USA. BMJ Open 2020 Nov 27;10(11):e041277. [doi: 10.1136/bmjopen-2020-041277] [Medline: 33247025]
13. Caffrey SL, Willoughby PJ, Pepe PE, Becker LB. Public use of automated external defibrillators. N Engl J Med 2002 Oct 17;347(16):1242-1247. [doi: 10.1056/NEJMoa020932] [Medline: 12393821]
14. Chen F, Li Y, Gong Y, Wei L, Wang J, Li Y. Evaluation of functional and electrical features of automatic external defibrillators in extreme altitude and temperature environments. Resusc Plus 2024 Mar;17:100562. [doi: 10.1016/j.resplu.2024.100562] [Medline: 38323138]
15. Jespersen SS, Kjoelbye JS, Christensen HC, et al. Functionality of registered automated external defibrillators. Resuscitation 2022 Jul;176:58-63. [doi: 10.1016/j.resuscitation.2022.05.013] [Medline: 35618078]
16. Olalere SO, Choi J. Controlled voltage of hot snare polypectomy device in electrosurgical device. J Eng (Stevenage) 2023 Dec 21;2023:1-14. [doi: 10.1155/2023/5521294]
17. Jonsson M, Berglund E, Müller MP. Automated external defibrillators and the link to first responder systems. Curr Opin Crit Care 2023 Dec 1;29(6):628-632. [doi: 10.1097/MCC.0000000000001109] [Medline: 37861209]

18. Sarkisian L, Mickley H, Schakow H, et al. Use and coverage of automated external defibrillators according to location in out-of-hospital cardiac arrest. Resuscitation 2021 May;162:112-119. [doi: 10.1016/j.resuscitation.2021.01.040] [Medline: 33581227]
19. Hanna DP, Erika B, Ellinor B, et al. Dispatcher nurses' experiences of handling drones equipped with automated external defibrillators in suspected out-of-hospital cardiac arrest - a qualitative study. Scand J Trauma Resusc Emerg Med 2024 Aug 21;32(1):74. [doi: 10.1186/s13049-024-01246-6] [Medline: 39169425]
20. Zègre-Hemsey JK, Cheskes S, Johnson AM, et al. Challenges & barriers for real-time integration of drones in emergency cardiac care: lessons from the United States, Sweden, & Canada. Resusc Plus 2024 Mar;17:100554. [doi: 10.1016/j.resplu.2024.100554] [Medline: 38317722]
21. Schierbeck S, Nord A, Svensson L, et al. Drone delivery of automated external defibrillators compared with ambulance arrival in real-life suspected out-of-hospital cardiac arrests: a prospective observational study in Sweden. Lancet Digit Health 2023 Dec;5(12):e862-e871. [doi: 10.1016/S2589-7500(23)00161-9] [Medline: 38000871]
22. Andelius L, Folke F. On-site bystanders or dispatched volunteer responders for bystander defibrillation: same goal, different paths. Resuscitation 2024 Feb;195:110105. [doi: 10.1016/j.resuscitation.2023.110105] [Medline: 38184179]
23. Karlsson L, Hansen CM, Vourakis C, et al. Improving bystander defibrillation in out-of-hospital cardiac arrests at home. Eur Heart J Acute Cardiovasc Care 2020 Nov;9(4_suppl):S74-S81. [doi: 10.1177/2048872619891675] [Medline: 32166951]
24. Lapidus O, Jonsson M, Svensson L, et al. Effects of a volunteer responder system for out-of-hospital cardiac arrest in areas of different population density - a retrospective cohort study. Resuscitation 2023 Oct;191:109921. [doi: 10.1016/j.resuscitation.2023.109921] [Medline: 37543160]

Abbreviations

AED: automated external defibrillator
CPR: cardiopulmonary resuscitation
EMS: emergency medical services
FEA: finite element analysis
OHCA: out-of-hospital cardiac arrest
PCB: printed circuit board

Edited by T Leung; submitted 29.09.23; peer-reviewed by A Akinfenwa, Anonymous; revised version received 02.03.25; accepted 01.07.25; published 19.08.25.

Please cite as:
Olalere SO
Effect of Thermal and Vibration Changes on Automated External Defibrillator Circuit Boards: Finite Element Analysis Study
JMIRx Med 2025;6:e53208
URL: https://xmed.jmir.org/2025/1/e53208
doi:10.2196/53208

Mothers' Knowledge of and Practices Toward Oral Hygiene of Children Aged 5-9 Years in Bangladesh: Cross-Sectional Study

Tahazid Tamannur[1], MPH; Sadhan Kumar Das[1], MPH; Arifatun Nesa[2], MPH; Foijun Nahar[1], MPH; Nadia Nowshin[1], MPH; Tasnim Haque Binty[1], MPH; Shafiul Azam Shakil[2], MPH; Shuvojit Kumar Kundu[3], MPH; Md Abu Bakkar Siddik[4], MPH; Shafkat Mahmud Rafsun[5], MPH; Umme Habiba[6], MPH; Zaki Farhana[7], MS; Hafiza Sultana[1], MPhil; Anton Abdulbasah Kamil[8], PhD; Mohammad Meshbahur Rahman[9], MS

[1]Department of Health Education, National Institute of Preventive and Social Medicine, Dhaka, Bangladesh

[2]Department of Public Health and Hospital Administration, National Institute of Preventive and Social Medicine, Mohakhali, Dhaka, Bangladesh

[3]Directorate General of Health Services, Ministry of Health & Family Welfare, Government of the People's Republic of Bangladesh, Dhaka, Bangladesh

[4]School of the Environment, Nanjing University, Nanjing, China

[5]Dental Speciality Center, Dhaka, Bangladesh

[6]BRAC James P Grant School of Public Health, BRAC University, Dhaka, Bangladesh

[7]Credit Information Bureau, Bangladesh Bank, Dhaka, Bangladesh

[8]Department of Business Administration, Istanbul Gelisim University, Istanbul, Turkey

[9]Department of Biostatistics, National Institute of Preventive and Social Medicine, Dhaka, Bangladesh

Corresponding Author:
Mohammad Meshbahur Rahman, MS
Department of Biostatistics, National Institute of Preventive and Social Medicine, Dhaka, Bangladesh

Related Articles:

Companion article: https://www.medrxiv.org/content/10.1101/2024.04.05.24305403v1

Companion article: https://med.jmirx.org/2025/1/e70142

Companion article: https://med.jmirx.org/2025/1/e70144

Companion article: https://med.jmirx.org/2025/1/e70145

Abstract

Background: Healthy oral hygiene is crucial for overall health and well-being. Parents' dental care knowledge and practices affect their children's oral health.

Objective: This study examined mothers' knowledge and practices regarding their children's oral hygiene through a cross-sectional survey.

Methods: This cross-sectional survey was conducted from January 1 to December 31, 2022, in Dhaka, Bangladesh. Mothers' knowledge and practices regarding their children's oral hygiene were assessed through a semistructured questionnaire. Statistical analyses, including the χ^2 test and Pearson correlation test, were performed. The Mann-Whitney *U* and Kruskal-Wallis 1-way ANOVA tests were also used to show the average variations in knowledge and practices among different sociodemographic groups.

Results: Of 400 participants, the mean age of mothers was 30.94 (SD 5.15) years, and 388 (97%) were of the Muslim faith, 347 (86.8%) were housewives, and 272 (68%) came from nuclear families. A total of 165 (41.3%) participants showed good knowledge of their children's oral hygiene, followed by 86 (21.5%) showing moderately average knowledge, 75 (18.8%) showing average knowledge, and 74 (18.5%) showing poor knowledge. A total of 182 (45.5%) mothers had children with good oral hygiene practices, followed by mothers with children who had average (n=78, 19.5%), moderately average (n=75, 18.8%), and poor (n=65, 16.3%) oral hygiene practices. The mother's knowledge level was significantly associated with age (*P*=.01), education (*P*<.001), family size (*P*=.03), and monthly income (*P*<.001). On the other hand, educational status (*P*=.002) and income (*P*=.04) were significantly associated with the mother's practices regarding their children's oral hygiene. Nonparametric analysis revealed that mothers who were older (mean knowledge score: 12.13, 95% CI 10.73-13.54 vs 11.21, 95% CI 10.85-11.58; *P*=.01), with a

bachelor's degree or higher (mean knowledge score: 12.93, 95% CI 12.55 - 13.31 vs 9.66, 95% CI 8.95 - 10.37; *P*<.001), who were working mothers (mean knowledge score: 12.30, 95% CI 11.72 - 12.89 vs 11.45, 95% CI 11.17 - 11.73; *P*=.03), and who had a higher family income (mean knowledge score: 12.49, 95% CI 12.0 - 12.98 vs 10.92, 95% CI 10.48 - 11.36; *P*<.001) demonstrated significantly higher levels of oral health knowledge. Conversely, good oral hygiene practices were significantly associated with higher maternal education (mean practice score: 6.88, 95% CI 6.54 - 7.22 vs 6.01, 95% CI 5.63 - 6.40; *P*<.001) and family income (mean practice score: 6.77, 95% CI 6.40 - 7.14 vs 5.96, 95% CI 5.68 - 6.24; *P*=.002). The mother's knowledge was also significantly and positively correlated (Pearson correlation coefficient *r*=0.301; *P*<.001) with their children's oral hygiene practices, shown by both the Pearson chi-square (χ^2=25.2; *P*<.001) test and correlation coefficient.

Conclusions: The mothers' knowledge and their children's oral hygiene practices were inadequate. The mother's age, education level, family size, and monthly income significantly influenced their knowledge level. Children's oral hygiene habits were significantly associated with family income and the mother's educational status. This underscores the need for educational programs, accessible dental care services, oral health education in the curriculum, media and technology involvement in oral health educational campaigns, and proper research and monitoring.

(JMIRx Med 2025;6:e59379) doi:10.2196/59379

KEYWORDS

mothers' knowledge and practices; oral hygiene; child oral health; Bangladesh

Introduction

According to the World Health Organization, dental caries, periodontal disease, tooth loss, mouth cancer, oro-dental trauma, noma, and congenital defects including cleft lip and palate are classified as oral diseases. Oral health issues are prevalent in low-income nations owing to poor socio–educational-economic circumstances [1]. In terms of general health and well-being, there is a significant connection between oral health and overall health [2,3]. It impacts individuals' capacity to do tasks, communicate, and engage in social interactions. Thus, it has an impact on both the physical and psychological aspects of an individual [4]. Most common oral health problems and conditions can be readily avoided by establishing suitable oral hygiene routines, such as twice daily brushing with the best toothbrush, using fluoride-containing toothpaste, and using the proper brushing technique [5]. Other preventive measures include eating a balanced diet low in free sugar, going to the dentist regularly for exams, and receiving treatment for illnesses when they are still in the early stages [6]. It can be minimized by practicing good oral hygiene habits, such as brushing and flossing teeth and visiting the dentist frequently [7].

Worldwide, over 2 billion individuals have dental caries in their permanent teeth, while 514 million children have dental caries in their primary teeth [8]. Early childhood caries (ECC) in children have been linked mostly to poor dental hygiene. Infants and toddlers with significant plaque accumulation were more likely to experience severe ECC and caries from birth to toddlerhood [9]. ECC has several causes, including excessive sugar intake, poor dental hygiene, inadequate fluoride exposure, and enamel abnormalities [10]. So, the development of caries and the acquisition of infection are substantially influenced by diet and feeding habits.

The children in Bangladesh have various infections and disorders [11-13]. Poor oral health is another prevalent health problem among them, which is still neglected [13]. As parents are the major caregivers, their involvement is crucial in the maintenance and development of excellent oral health in children, such as teaching healthy eating and drinking habits [14]. In addition, several factors impact the dental health of children, including the mother's level of education, the mother's work situation, and her understanding of oral hygiene [15]. The adoption of good oral health practices in children is influenced by the parents', and particularly the mother's, oral health knowledge, attitudes, and awareness [16]. An Indian study found that the oral hygiene quality of children aged 12 years was shown to be significantly influenced by their mother's oral hygiene knowledge [17]. Children with high rates of dental caries and low rates of fillings were found to have parents with inadequate oral health literacy, according to another study [18]. As a result, it is essential for parents, and particularly mothers, to have awareness about oral health. Scholars argued that a mother's knowledge about oral health and the consequence of adequate dental hygiene has a beneficial impact on their children's dental well-being and adherence to dental care practices [19,20].

Research on dental caries awareness among parents in Pakistan has found low levels of knowledge about oral hygiene standards [21]. A study conducted in India on the oral health status of children aged 3 - 6 years and their mother's oral health–related knowledge, attitude, and practices found most mothers had a medium level of knowledge, an average level of attitude, and a high level of practices regarding oral health [22]. Another study in Malaysia on parental knowledge and practices in preschool children's oral health found that the majority had good knowledge [23]. Numerous studies have been conducted globally regarding parents' or mothers' oral hygiene knowledge and practices, but these have been insufficient, particularly among mothers in Bangladesh. There is a lack of research investigating the extent to which mothers are aware of and follow oral hygiene practices. Hence, this study aimed to assess mothers' level of oral hygiene knowledge and practices regarding their 5- to 9-year-old children.

Methods

This study followed the STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) guidelines to prepare the manuscript, and the STROBE checklist is provided in Multimedia Appendix 1.

Ethical Considerations

Permission to conduct this study was given by the institutional review board of the National Institute of Preventive and Social Medicine (NIPSOM), Bangladesh (Ref NIPSOM/IRB/2017/09). The Shaheed Suhrawardy Medical College Hospital and Dhaka Dental College Hospital provided the necessary documentation. Both written and verbal consent were taken before initiating the interview. Participants received an overview of the study's goals, and those who consented were eventually included. No compensation was given to the participants, and data anonymity was strictly maintained.

Study Setting and Participants

This cross-sectional study was conducted from January 1 to December 31, 2022, in two tertiary-level hospitals in Dhaka South named Shaheed Suhrawardy Medical College Hospital and Dhaka Dental College Hospital in Dhaka City. Mothers of children aged 5-9 years who visited these tertiary hospitals were interviewed through a semistructured questionnaire.

Study Pretesting

To observe the overall scenario including questionnaire information, possible sampling techniques, and approximate nonresponse rate in the study, we first performed a pretest of the study. The pretesting was conducted among 50 mothers of children aged 5 - 9 years in the Sapporo Dental College & Hospital located at Dhaka North.

Sampling Technique and Sample Size

A convenience sampling technique was followed for this study. During the literature search, no study was found that assessed the knowledge and practices toward children's oral hygiene among Bangladeshi mothers. However, a study was found from India with a similar sociodemography. Mohandass et al [20] showed that the prevalence of adequate knowledge and practices were 58% and 57%, respectively. The sample size for this study was calculated using the below equation.

$$(1) n=z2pqd2$$

The sample size when P=.58 for the mother's knowledge was:

$$n=1.962\times0.58\times(1-0.58)0.052=375$$

Similarly, the sample size when P=.57 for the mother's practice level was:

$$n=1.962\times0.57\times(1-0.57)0.052=377$$

Therefore, we initially chose a maximum of 377 as the required sample size. Considering a maximum 5% nonresponse rate (based on pretesting), we rounded up this figure and selected 400 as an approximate sample size for the study.

Selection Criteria

The inclusion criteria for this study were mothers of Bangladeshi nationality who were living in Dhaka for at least 1 year, mothers of children aged 5 - 9 years, and mothers who provided consent and agreed to participate in the study. The exclusion criteria for the study were mothers who were not Bangladeshi but currently living in Dhaka, mothers of children older than 10 years or younger than 5 years, and mothers younger than 21 years or older than 48 years.

Sociodemographic Variables

Respondents' sociodemographic variables such as age (21-48 years), religion (Muslim, non-Muslim), educational status (up to primary, secondary, higher secondary, and bachelor's degree or higher), occupational status (housewife, working), family type (nuclear, joint), family size (<5 persons, ≥5 persons), and monthly family income (≤20,000 BDT, 20,001 - 40,000 BDT, ≥40,001 BDT; a currency exchange rate of 101.85 BDT=US $1 was used) were the independent variables in this study.

Measurement of Knowledge and Practice

The study used 15 variables to assess mothers' knowledge and 13 to assess their children's practices related to oral hygiene (Multimedia Appendices 2 and 3). Both knowledge and practice questions were adopted from the existing literature and revised according to our selection criteria. The summation scoring technique was used in computation, and the descriptive statistics, including percentiles, were observed. The range for the knowledge and practice scores were 1-15 and 1-13, respectively. According to the percentile approach, knowledge was classified into four levels: poor (<25% percentile cut point: ≤9.999), moderately average (25% - 49% percentile cut point: 10.0 - 11.99), average (50% - 74% percentile cut point: 12.0 - 12.99), and good knowledge (≥75% percentile cut point: ≥13.0) [24]. Practices were also classified into four levels: poor (<25% percentile cut point:≤4.99), moderately average (25% - 49% percentile cut point: 5.0 - 5.999), average (50% - 74% percentile cut point: 6.0 - 6.99), and good practices (≥75% percentile cut point: ≥7.0). For all cases, the cut points were statistically evident [25,26].

Data Quality Control

To ensure the reliability and validity of the study findings, we observed the reliability analysis for both knowledge and practice variables, yielding a Cronbach α; the reliability coefficient values for the variables related to knowledge and practice were found to be 0.78 and 0.81, indicating acceptable internal consistency.

Statistical Analysis

Descriptive statistics were performed to present participants' sociodemographic characteristics and mean knowledge and practice scores. The Pearson χ^2 test and Pearson correlation coefficient were used as a bivariate analysis. Since both knowledge and practice scores did not follow normality, we performed the Mann-Whitney U test and Kruskal-Wallis 1-way ANOVA test to show the mean knowledge and practice score variations between two (eg, housewife vs working mother) and more than two groups (eg, different age groups), respectively.

Necessary assumptions were checked before performing the statistical analysis. All the data management and statistical analyses were carried out through SPSS Statistics 27.0 (IBM Corp). The *P* value was observed for all the cases at a 5% level, and 95% was considered as the CI [27-29].

Results

Sociodemographic Characteristics of the Respondents

The majority of the respondents (n=209, 52.3%) were within the 21 - 30 years age group, followed by 44% (n=176) in the 31 - 40 years age group. Most (n=57, 39.3%) respondents had a secondary level of education. Most were Muslims (n=388, 97%) and housewives (n=347, 86.8%). Many of the respondents (n=157, 39.3%) had a monthly family income of 20,001 - 40,000 BDT (US $206.19-$392.73) per month. About 13.3% (n=53) of mothers were working (Table 1).

Table . Distribution of sociodemographic characteristics of the respondents (N=400).

Characteristics	Respondents, n (%)
Age group (years)	
21 - 30	209 (52.3)
31 - 40	176 (44.0)
41 - 48	15 (3.8)
Religion	
Muslim	388 (97.0)
Non-Muslim	12 (3.0)
Educational status	
Up to primary	76 (19.0)
Secondary	157 (39.3)
Higher secondary	68 (17.0)
Bachelor's degree or higher	99 (24.8)
Occupation	
Housewife	347 (86.8)
Working	53 (13.3)
Family type	
Nuclear	272 (68.0)
Joint	128 (32.0)
Number of family members	
<5 persons	193 (48.3)
≥5 persons	207 (51.8)
Monthly family income (BDT)[a]	
≤20,000	143 (35.8)
20,001 - 40,000	157 (39.3)
≥40,001	100 (25.0)

[a]A currency exchange rate of 101.85 BDT=US $1 was used.

Knowledge Among Mothers Regarding Their Children's Oral Hygiene

Multimedia Appendix 4 shows the mothers' knowledge scores regarding their children's oral hygiene. Among the 400 mothers, more than 90% (n=360) knew the importance of brushing teeth, while 82.3% (n=329) and 80.8% (n=323) knew the recommended frequency and appropriate time for brushing teeth, respectively. Surprisingly, only 29.5% (n=118) and 38.5% (n=154) knew the duration for brushing teeth and that fluoride protects against caries, respectively. However, most of the respondents knew about the "importance of cleaning tongue" (n=365, 91.3%), "gingival disease common cause of gum bleeding" (n=286, 71.5%), "brushing and flossing protect against bleeding gum" (n=243, 60.8%), "yellow coating plaque" (n=362, 90.5%), "sugary item cause caries" (n=387, 96.8%), "soft drinks cause caries" (n=295, 73.8%), and "regular brushing protects against caries" (n=380, 95%).

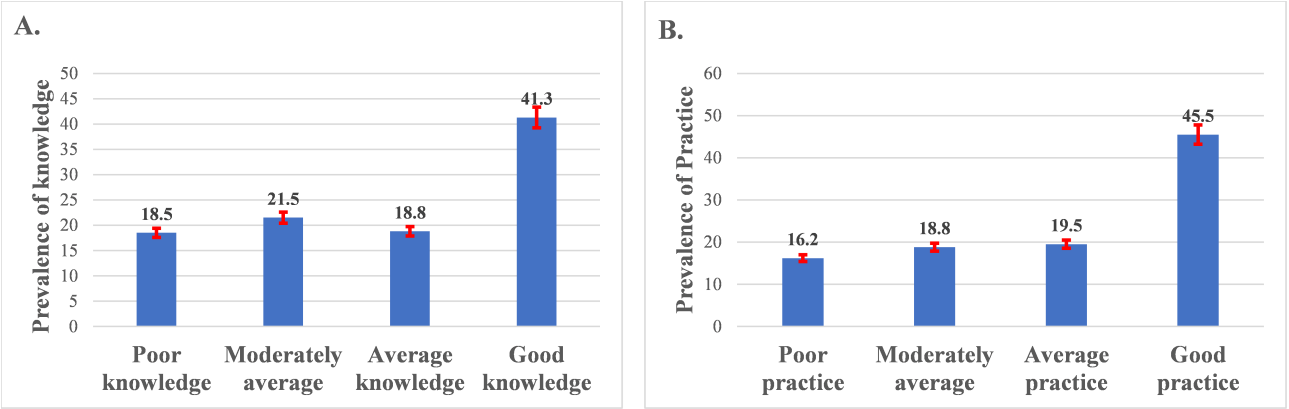
Mothers' Practices Regarding Their Children's Oral Hygiene

Multimedia Appendix 5 shows the individual distribution of mothers' practices regarding their children's oral hygiene. Most (n=381, 95.3%) of the mothers reported that their child brushed their teeth regularly, 99% (n=396) of children used a toothbrush, 62% (n=248) changed their toothbrush every 3 - 4 months or if the bristles were frayed, 97.8% (n=391) used their toothpaste, and 77.8% (n=311) rinsed their mouth after eating. Surprisingly, 44.3% (n=177) of children brushed their teeth twice daily, 42% (n=168) cleaned their tongues, and 2.8% (n=11) used floss. Only 12.5% (n=50) were given sugary items with meals, and 0.3% (n=1) were taken to dentists every 6 months.

Overall Knowledge and Practice Levels of the Respondents

Figure 1 depicts the level of knowledge and practices of mothers regarding their children's oral hygiene and the association with the mother's educational status. Only 41.3% (n=165) had good knowledge, while 18.5% (n=74) had poor knowledge (Figure 1A). Similarly, only 45.5% (n=182) of the mothers showed good practices, while 16.2% (n=65) showed poor practice levels (Figure 1B).

Figure 1. (A) Distribution of the overall knowledge of mothers. (B) Distribution of the overall practices of mothers.



Sociodemographic Variations in the Mother's Knowledge Level Regarding Their Children's Oral Hygiene

A total of 66.7% (10/15) of mothers aged 41-48 years had good knowledge regarding their children's oral hygiene. The Pearson χ^2 association test revealed that mothers' knowledge levels were significantly associated with age (P=.01), education (P<.001), family size (P=.03), and monthly income (P<.001; Table 2).

Table . Association of mothers' knowledge with sociodemographic characteristics.

Characteristics	Poor knowledge, n (%)	Moderately average, n (%)	Average knowledge, n (%)	Good knowledge, n (%)	*P* value[a]
Age group (years)					.01
21 - 30 (n=209)	43 (20.6)	55 (26.3)	39 (18.7)	72 (34.4)	
31 - 40 (n=176)	28 (15.9)	29 (16.5)	36 (20.5)	83 (47.2)	
41 - 48 (n=15)	3 (20.0)	2 (13.3)	0 (0.0)	10 (66.7)	
Religion					.22
Muslim (n=388)	72 (18.6)	86 (22.2)	72 (18.6)	158 (40.7)	
Non-Muslim (n=12)	2 (16.7)	0 (0.0)	3 (25.0)	7 (58.3)	
Educational status					<.001
Up to primary (n=76)	32 (42.1)	22 (28.9)	8 (10.5)	14 (18.4)	
Secondary (n=157)	29 (18.5)	46 (29.3)	33 (21.0)	49 (31.2)	
Higher secondary (n=68)	8 (11.8)	8 (11.8)	16 (23.5)	36 (52.9)	
Bachelor's degree or higher (n=99)	5 (5.1)	10 (10.1)	18 (18.2)	66 (66.7)	
Occupation					.10
Housewife (n=347)	68 (19.6)	77 (22.2)	67 (19.3)	135 (38.9)	
Working (n=53)	6 (11.3)	9 (17.0)	8 (15.1)	30 (56.6)	
Family type					.06
Nuclear (n=272)	46 (16.9)	52 (19.1)	59 (21.7)	115 (42.3)	
Joint (n=128)	28 (21.9)	34 (26.6)	16 (12.5)	50 (39.1)	
Number of family members					.03
<5 persons (n=193)	27 (14.0)	36 (18.7)	39 (20.2)	91 (47.2)	
≥5 persons (n=207)	47 (22.7)	50 (24.2)	36 (17.4)	74 (35.7)	
Monthly family income (BDT)[b]					<.001
≤20,000 (n=143)	38 (26.6)	37 (25.9)	25 (17.5)	43 (30.1)	
20,001 - 40,000 (n=157)	26 (16.6)	33 (21.0)	37 (23.6)	61 (38.9)	
≥41,001 (n=100)	10 (10.0)	16 (16.0)	13 (13.0)	61 (61.0)	

[a]χ^2/Fisher exact test.

[b]A currency exchange rate of 101.85 BDT=US $1 was used.

Sociodemographic Variation of the Mother's Practice Level Regarding Their Children's Oral Hygiene

Table 3 represents the association between mothers' sociodemographic characteristics and their practices regarding their children's oral hygiene. The analysis found that more than half (n=8, 53.3%) of older-aged mothers had good practices, and 66.7% (n=60) of mothers with a bachelor's degree or higher showed good practices regarding their children's oral hygiene. The educational status (*P*=.002) and income (*P*=.04) were significantly associated with the mothers' practices regarding their children's oral hygiene (Table 3).

Table . Association between sociodemographic characteristics and practice level regarding their children's oral hygiene.

Characteristics	Poor practice, n (%)	Moderately average, n (%)	Average practice, n (%)	Good practice, n (%)	*P* value[a]
Age group (years)					.34
21 - 30 (n=209)	34 (16.3)	44 (21.1)	46 (22.0)	85 (40.7)	
31 - 40 (n=176)	30 (17.0)	27 (15.3)	30 (17.0)	89 (50.6)	
41 - 48 (n=15)	1 (6.7)	4 (26.7)	2 (13.3)	8 (53.3)	
Religion of the respondents					.42
Muslim (n=388)	65 (16.8)	73 (18.8)	76 (19.6)	174 (44.8)	
Non-Muslim (n=12)	0 (0.0)	2 (16.7)	2 (16.7)	8 (66.7)	
Educational status of the respondent					.002
Up to primary (n=76)	15 (19.7)	19 (25.0)	6 (7.9)	36 (47.4)	
Secondary (n=157)	27 (17.2)	34 (21.7)	41 (26.1)	55 (35.0)	
Higher secondary (n=68)	12 (17.6)	12 (17.6)	13 (19.1)	31 (45.6)	
Bachelor's degree or higher (n=99)	11 (11.1)	10 (10.1)	18 (18.2)	60 (60.6)	
Occupation of the respondent					.24
Housewife (n=347)	60 (17.3)	68 (19.6)	65 (18.7)	154 (44.4)	
Working (n=53)	5 (9.4)	7 (13.2)	13 (24.5)	28 (52.8)	
Family type of the respondent					.98
Nuclear (n=272)	43 (15.8)	51 (18.8)	54 (19.9)	124 (45.6)	
Joint (n=128)	22 (17.2)	24 (18.8)	24 (18.8)	58 (45.3)	
Number of family members					.93
<5 persons (n=193)	30 (15.5)	38 (19.7)	36 (18.7)	89 (46.1)	
≥5 persons (n=207)	35 (16.9)	37 (17.9)	42 (20.3)	93 (44.9)	
Monthly family income (BDT)[b]					.04
≤20,000 (n=143)	30 (21.0)	30 (21.0)	28 (19.6)	55 (38.5)	
20,001 - 40,000 (n=157)	22 (14.0)	31 (19.7)	35 (22.3)	69 (43.9)	
≥41,001 (n=100)	13 (13.0)	14 (14.0)	15 (15.0)	58 (58.0)	

[a]χ^2/Fisher exact test significant level.

[b]A currency exchange rate of 101.85 BDT=US $1 is applicable.

Variation in Knowledge and Practices of the Respondents

A significant difference in respondents' knowledge and practices with sociodemographic characteristics was observed (Table 4). The analysis found that the knowledge was comparatively higher among mothers of higher age groups compared to lower age groups (mean knowledge score: 12.13, 95% CI 10.73-13.54 vs 11.23, 95% CI 10.85-11.58; *P*=.01). Similarly, both the knowledge and practice behaviors were significantly higher among mothers with higher education and income than their counterparts. In addition, working mothers and mothers with small families had significantly higher knowledge (Table 4).

Table . Knowledge and practice variation of mothers according to sociodemographic characteristics.

Characteristics	Knowledge score (range 1-15), mean (95% CI)	*P* value[a]	Practice score (range 1-13), mean (95% CI)	*P* value[a]
Age group (years)		.01		.21
21 - 30 (n=209)	11.21 (10.85 - 11.58)		6.13 (5.92 - 6.35)	
31 - 40 (n=176)	11.93 (11.56 - 12.29)		6.36 (6.09 - 6.62)	
41 - 48 (n=15)	12.13 (10.73 - 13.54)		6.8 (5.67 - 7.93)	
Religion		.22		.19
Muslim (n=388)	11.54 (11.28 - 11.80)		6.24 (6.07 - 6.41)	
Non-Muslim (n=12)	12.25 (10.53 - 13.97)		6.83 (6.08 - 7.59)	
Educational status		<.001		<.001
Up to primary (n=76)	9.66 (8.95 - 10.37)		6.01 (5.63 - 6.40)	
Secondary (n=157)	11.32 (10.97 - 11.67)		6.01 (5.75 - 6.27)	
Higher secondary (n=68)	12.26 (11.71 - 12.82)		6.19 (5.79 - 6.59)	
Bachelor's degree or higher (n=99)	12.93 (12.55 - 13.31)		6.88 (6.54 - 7.22)	
Occupation		.03		.12
Housewife (n=347)	11.45 (11.17 - 11.73)		6.21 (6.02 - 6.39)	
Working (n=53)	12.30 (11.72 - 12.89)		6.59 (6.19 - 6.98)	
Family type		.13		.88
Nuclear (n=272)	11.7 (11.39 - 12.00)		6.25 (6.05 - 6.45)	
Joint (n=128)	11.28 (10.81 - 11.75)		6.28 (5.98 - 6.59)	
Number of family members		<.001		.95
<5 persons (n=193)	11.96 (11.6 - 12.32)		6.27 (6.03 - 6.51)	
≥5 persons (n=207)	11.19 (10.84 - 11.55)		6.25 (6.01 - 6.48)	
Monthly family income (BDT)[b]		<.001		.002
≤20,000 (n=143)	10.92 (10.48 - 11.36)		5.96 (5.68 - 6.24)	
20,001 - 40,000 (n=157)	11.56 (11.17 - 11.95)		6.20 (5.96 - 6.45)	
≥40,001 (n=100)	12.49 (12.0 - 12.98)		6.77 (6.40 - 7.14)	

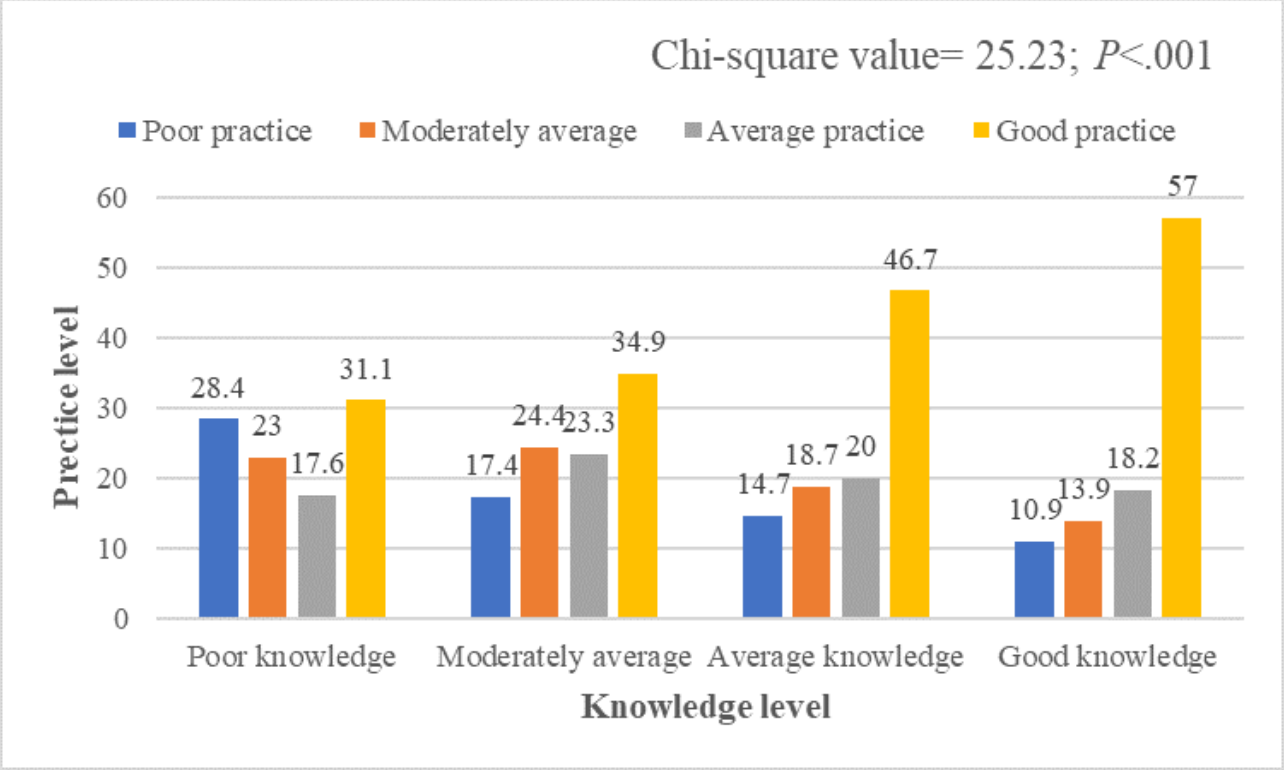
[a]Mann-Whitney *U* test and Kruskal-Wallis 1-way ANOVA test.

[b]A currency exchange rate of 101.85 BDT=US $1 is applicable.

Association Between Mothers' Oral Hygiene Knowledge and Practice Levels

Figure 2 represents the association between mothers' oral hygiene knowledge and practice levels. Over 50% of mothers with good knowledge had good practice behaviors regarding their children's oral hygiene. The Pearson correlation coefficient analysis also found a significant and positive association (r=0.301; P<.001) between the knowledge and practice scores of the respondents (Multimedia Appendix 6).

Figure 2. Association between mothers' knowledge level and practice level.



Discussion

Principal Findings

Oral health is an integral component of overall health, and it is important in our everyday lives. This study intended to evaluate mothers' knowledge and practices regarding their children's oral hygiene. An increased knowledge level was observed among older mothers, those with higher education levels, working mothers, and mothers from higher income groups. Similarly, good practices regarding children's oral hygiene were associated with the mother's education level and economic status.

Comparison to Prior Work

To maintain oral health, brushing twice a day is standard [30]. The study found that most mothers know the standard brushing recommendation for their children. Many mothers also agreed that gingival disease was the most common cause of gum bleeding, and brushing and flossing could protect against bleeding gums. The findings align with the existing literature [31]. If one wants to protect themselves against any kind of dental sickness, brushing regularly is required [32]. Over 50% of the mothers in our study agreed with this statement, which is comparable to existing research findings [33]. In this study, less than half of the mothers had good knowledge regarding their children's oral hygiene, and nearly 1 in every 5 mothers had poor knowledge. The findings suggest that health education programs among mothers regarding their children's oral hygiene are needed. Various education and awareness programs, including television, social and mass media campaigns, and community-based educational interventions may improve mothers' knowledge regarding children's oral hygiene [34-36].

In this study, the mother's knowledge regarding their children's oral hygiene was significantly associated with their age, and mothers in the higher age group had comparatively higher knowledge than those in the lower age group. The finding is comparable to many studies that suggest oral health educational programs for younger mothers [34,37,38]. The mother's educational status and monthly family income were two important predictors for increasing their children's oral hygiene knowledge and practices. Parents with higher education were more aware of their children's dental health [39,40]. Our research results align with the existing literature that indicates that mothers who have attained a university degree possess superior knowledge about oral health in comparison to those with a lower level of education [41]. This might be rationalized by the deduction that women with a lower level of education may lack awareness about the consequences of probable risk factors linked to the progression of oral disorders. Consequently, health awareness and promotion play a vital role for mothers who have inadequate educational backgrounds [40,42,43]. Our results align with the existing research, which demonstrates that mothers with extensive knowledge tend to promote good oral health habits in their children [25].

Strengths and Limitations

This study aimed to identify the variables that impact oral hygiene habits among mothers and evaluate their level of knowledge and compliance with oral hygiene practices. The primary merit of this study is the results. We identified the variables that influence individuals' understanding and behaviors related to oral hygiene. We experienced a few limitations during this study. First, this was cross-sectional research, which lacks strength in cause-effect analysis. Second, the study was conducted among mothers visiting tertiary-level hospitals in

Dhaka. Therefore, there is a chance of nonresponse bias due to convenience sampling.

Future Directions

Maintaining good oral hygiene is crucial for every child's overall health; mothers, in particular, play a vital role in this regard. Based on our study findings, the following recommendations may help enhance maternal knowledge and improve children's oral hygiene practices.

Educational Workshops and School-Based Initiatives

Community-based educational programs including workshops and seminars may help educate mothers of different age groups [34,40]. These workshops should focus on the importance of oral hygiene, practical tips for maintaining children's oral health, and common misconceptions. Monthly informational sessions on oral hygiene practices facilitated by dental health professionals and community health centers could play an important role in improving children's oral hygiene practices. Various school-based initiatives, like partnering with schools to offer regular seminars and distributing informative materials to parents during parent-teacher meetings that emphasize the critical role of oral hygiene from an early age, could be implemented [37].

Incorporate Oral Health Education Into the Curriculum

Integration of basic oral health education into the curriculum of early childhood education programs, ensuring that children learn about oral hygiene from a young age, may help children improve their oral hygiene practices [39,44]. Various programs within schools that encourage parental involvement in learning about and practicing good oral hygiene, and providing resources and support for mothers to reinforce these practices at home may help children improve their oral hygiene practices [44].

Media and Technology Use

Launching social media campaigns targeting mothers; using platforms like Facebook, Instagram, and YouTube to disseminate information on children's oral hygiene; and featuring engaging content such as infographics, videos, and interactive question-and-answer sessions with dental professionals could also be influential initiatives [35].

Research and Monitoring

Support should also be provided for ongoing research to monitor the effectiveness of these initiatives and to identify new trends and needs related to children's oral hygiene [45]. Establishing feedback mechanisms, such as surveys and focus groups, can help gather insights from mothers on the effectiveness of current programs and identify areas for improvement.

Conclusion

This study revealed that mothers' knowledge and practices regarding their children's oral health were insufficient. The mother's age, education level, family size, and monthly income significantly influenced their knowledge level. Children's oral hygiene habits were significantly associated with family income and the mother's educational status. Women aged 41-48 years with a bachelor's degree or higher, from higher socioeconomic backgrounds, and with school-aged children demonstrated significantly higher levels of knowledge. Mothers with higher socioeconomic status and more education demonstrated a much higher level of dental hygiene practices for their children. The mother's knowledge regarding their children's oral hygiene had positive effects and significantly correlated with their children's oral hygiene practices. The findings of this study emphasize the need for educational and school-based initiatives, accessible dental care services, oral health education in the curriculum, media and technology involvement in oral health educational campaigns, and proper research and monitoring.

Acknowledgments

We acknowledge the Department of Biostatistics, National Institute of Preventive and Social Medicine for their technical support during the study. We are also grateful to all participants included in this study. We are thankful to the Supporo Dental College and Hospital for their administrative support during data collection.

Data Availability

The datasets generated or analyzed during this study were deposited onto figshare [46].

Authors' Contributions

Conceptualization: TT, MMR, HS
Formal analysis: TT, MMR
Investigation: TT, SKD, AN, FN, NN, THB, SAS, SKK, MABS, SMR, UH, ZF, MMR
Methodology: TT, MMR
Project administration: TT, MMR
Supervision: MMR
Funding acquisition: TT, SKD, FN, NN, THB, SAS, SKK, SMR, UH, ZF, AAK, MMR
Validation: HS, AAK, MMR
Visualization: TT, SKD, MMR
Writing - original draft: TT, SKD, MMR
Investigation: TT, SKD, AN, FN, NN, THB, SAS, SKK, MABS, SMR, UH, ZF, HS, AAK, MMR

Writing - review and editing: TT, SKD, AN, FN, NN, THB, SAS, SKK, MABS, SMR, UH, ZF, HS, AAK, MMR

Conflicts of Interest

None declared.

Multimedia Appendix 1
STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) statement: checklist of items that should be included in reports of cross-sectional studies.
[PDF File, 76 KB - xmed_v6i1e59379_app1.pdf]

Multimedia Appendix 2
List of variables used to assess mothers' knowledge regarding their children's oral hygiene.
[DOCX File, 14 KB - xmed_v6i1e59379_app2.docx]

Multimedia Appendix 3
List of variables used to assess mothers' practices regarding their children's oral hygiene.
[DOCX File, 14 KB - xmed_v6i1e59379_app3.docx]

Multimedia Appendix 4
Mothers' individual knowledge regarding their children's oral hygiene.
[DOCX File, 14 KB - xmed_v6i1e59379_app4.docx]

Multimedia Appendix 5
Mothers' individual practices regarding their children's oral hygiene.
[DOCX File, 14 KB - xmed_v6i1e59379_app5.docx]

Multimedia Appendix 6
Correlation between knowledge and practice scores.
[DOCX File, 13 KB - xmed_v6i1e59379_app6.docx]

References

1. Global oral health status report: towards universal health coverage for oral health by 2030. World Health Organization. 2022 Nov 18. URL: https://www.who.int/publications/i/item/9789240061484 [accessed 2025-01-13]
2. Alama MA, Abdullah US, Mofiz M, Aktar S, Karim M. Oral health status among the under five children attending at OPD of Dhaka Dental College Hospital. Update Dent Coll J 2016 Apr 7;5(2):9-17. [doi: 10.3329/updcj.v5i2.27263]
3. Haque MF, Rahman MM, Alif SM, et al. Estimation and prediction of doubling time for COVID-19 epidemic in Bangladesh: a study of first 14 month's daily confirmed new cases and deaths. Global Biosecurity 2021 Apr 26;3(1). [doi: 10.31646/gbio.91]
4. Shah PM, Jeevanadan G. Oral hygiene maintenance in children - a survey on parental awareness. Int J Pharm Res 2017;9(3). [doi: 10.31838/ijpr/2020.12.01.311]
5. Toothbrushes. American Dental Association. 2019. URL: https://www.ada.org/resources/research/science-and-research-institute/oral-health-topics/toothbrushes [accessed 2022-11-27]
6. Lawal FB, Fagbule OF, Akinloye SJ, Lawal TA, Oke GA. Impact of oral hygiene habits on oral health-related quality of life of in-school adolescents in Ibadan, Nigeria. Front Oral Health 2022 Sep 9;3:979674. [doi: 10.3389/froh.2022.979674] [Medline: 36338573]
7. Saadaldina SA, Eldwakhly E, Alnazzawi AA, et al. Awareness and practice of oral health measures in Medina, Saudi Arabia: an observational study. Int J Environ Res Public Health 2020 Dec 6;17(23):1-10. [doi: 10.3390/ijerph17239112] [Medline: 33291281]
8. Oral health. World Health Organization. 2023. URL: https://www.who.int/health-topics/oral-health#tab=tab_1 [accessed 2025-01-20]
9. Duangthip D, Chu CH. Challenges in oral hygiene and oral health policy. Front Oral Health 2020 Oct 7;1:575428. [doi: 10.3389/froh.2020.575428] [Medline: 35047981]
10. Anil S, Anand PS. Early childhood caries: prevalence, risk factors, and prevention. Front Pediatr 2017 Jul 18;5:157. [doi: 10.3389/fped.2017.00157] [Medline: 28770188]
11. Hossain MS, Tasnim S, Chowdhury MA, Chowdhury FIF, Hossain D, Rahman MM. Under‐five children's acute respiratory infection dropped significantly in Bangladesh: an evidence from Bangladesh demographic and health survey, 1996–2018. Acta Paediatr 2022 Oct;111(10):1981-1994. [doi: 10.1111/apa.16447] [Medline: 35678484]

12. Islam GMR, Rahman MM, Hasan MI, Tadesse AW, Hamadani JD, Hamer DH. Hair, serum and urine chromium levels in children with cognitive defects: a systematic review and meta-analysis of case control studies. Chemosphere 2022 Mar;291(Pt 2):133017. [doi: 10.1016/j.chemosphere.2021.133017] [Medline: 34813844]
13. Sony SA, Haseen F, Islam SS, Chowdhury SF. Knowledge and practice of oral health and hygiene and oral health status among school going adolescents in a rural area of Sylhet District. Community Based Med J 2022 Jan 10;10(1):30-36. [doi: 10.3329/cbmj.v10i1.58642]
14. Mohammadi TM, Hajizamani A, Bozorgmehr E. Improving oral health status of preschool children using motivational interviewing method. Dent Res J (Isfahan) 2015;12(5):476-481. [doi: 10.4103/1735-3327.166231] [Medline: 26604963]
15. Holm AK. Caries in the preschool child: international trends. J Dent 1990 Dec;18(6):291-295. [doi: 10.1016/0300-5712(90)90125-x] [Medline: 2074302]
16. Al-Batayneh OB, Al-Khateeb HO, Ibrahim WM, Khader YS. Parental knowledge and acceptance of different treatment options for primary teeth provided by dental practitioners. Front Public Health 2019 Nov 7;7:322. [doi: 10.3389/fpubh.2019.00322] [Medline: 31788466]
17. Chand S, Chand S, Dhanker K, Chaudhary A. Impact of mothers′ oral hygiene knowledge and practice on oral hygiene status of their 12-year-old children: a cross-sectional study. J Indian Assoc Public Health Dent 2014;12(4):323-329. [doi: 10.4103/2319-5932.147681]
18. Khodadadi E, Niknahad A, Sistani MMN, Motallebnejad M. Parents' oral health literacy and its impact on their children's dental health status. Electron Physician 2016 Dec 25;8(12):3421-3425. [doi: 10.19082/3421] [Medline: 28163858]
19. Al-Zahrani AM, Al-Mushayt AS, Otaibi MF, Wyne AH. Knowledge and attitude of Saudi mothers towards their preschool children's oral health. Pak J Med Sci 2014 Jul;30(4):720-724. [doi: 10.12669/pjms.304.5069] [Medline: 25097504]
20. Mohandass B, Chaudhary H, Pal GK, Kaur S. Knowledge and practice of rural mothers on oral hygiene for children. Indian J Continuing Nurs Education 2021;22(1):39-43. [doi: 10.4103/IJCN.IJCN_7_20]
21. Ali Leghari M, Tanwir F, Ali H. Association of dental caries and parents knowledge of oral health, a cross-sectional survey of schools of Karachi, Pakistan. J Pakistan Dent Assoc 2018 May 15;23(1) [FREE Full text]
22. Bennadi D, Reddy CVK, Sunitha S, Kshetrimayum N. Oral health status of 3-6 year old children and their mother's oral health related knowledge, attitude and practices in Mysore City, India. Asian J Med Sci 2014 Sep 15;6(2):66-71. [doi: 10.3126/ajms.v6i2.11097]
23. Bakar N, Mamat Z. Parental knowledge and practices on preschool children oral healthcare in Nibong Tebal Penang, Malaysia. JOJ Nurs Health Care 2018 Apr;7(4). [doi: 10.19080/JOJNHC.2018.07.555716]
24. Kabar AME, Elzahaf RA, Shakhatreh FM. The relationship between oral health knowledge mothers and dental caries in Tripoli, Libya. Saudi J Oral Dent Res 2019 Jul 22;4(7):463-467. [doi: 10.21276/sjodr.2019.4.7.7]
25. Alzaidi SS, Alanazi IA, Abo Nawas OM, Mulla MA. Childhood oral health: maternal knowledge and practice in Tabuk, Saudi Arabia. Egyptian J Hosp Med 2018 Jan;70(9):1544-1551. [doi: 10.12816/0044681]
26. Das SK, Tamannur T, Nesa A, et al. Exploring the knowledge and practices on road safety measures among motorbikers in Dhaka, Bangladesh: a cross-sectional study. Inj Prev 2023 Nov 28:ip-2023-045071. [doi: 10.1136/ip-2023-045071] [Medline: 38050086]
27. Nachar N. The Mann-Whitney U: a test for assessing whether two independent samples come from the same distribution. Tutorials Quant Methods Psychol 2008;4(1):13-20. [doi: 10.20982/tqmp.04.1.p013]
28. Kim HY. Statistical notes for clinical researchers: nonparametric statistical methods: 2. Nonparametric methods for comparing three or more groups and repeated measures. Restor Dent Endod 2014 Nov;39(4):329-332. [doi: 10.5395/rde.2014.39.4.329] [Medline: 25383354]
29. Rahman MM, Hamiduzzaman M, Akter MS, et al. Frailty indexed classification of Bangladeshi older adults' physio-psychosocial health and associated risk factors- a cross-sectional survey study. BMC Geriatr 2021 Jan 6;21(1):3. [doi: 10.1186/s12877-020-01970-5] [Medline: 33402094]
30. Manzoor F, Iqbal Z, Ahmed K, Khayyam U, Malhi P, Khalid M. Assessment of parental knowledge and attitude regarding oral health status of their children in District Mirpurkhas Sindh, Pakistan. Pakistan J Med Health Sci 2021 Apr;15(4):1352-1355 [FREE Full text]
31. Salama AA, Konsowa EM, Alkalash SH. Mothers' knowledge, attitude, and practice regarding their primary school children's oral hygiene. Menoufia Med J 2020 Mar 25;33(1):11-17. [doi: 10.4103/mmj.mmj_300_19]
32. Tonetti MS, Bottenberg P, Conrads G, et al. Dental caries and periodontal diseases in the ageing population: call to action to protect and enhance oral health and well-being as an essential component of healthy ageing - consensus report of group 4 of the joint EFP/ORCA workshop on the boundaries between caries and periodontal diseases. J Clin Periodontol 2017 Mar;44 Suppl 18:S135-S144. [doi: 10.1111/jcpe.12681] [Medline: 28266112]
33. Ibrahim R, Helaly MO, Ahmed EMA. Assessment of brushing techniques in school children and its association with dental caries, Omdurman, 2019. Int J Dent 2021 Jan 21;2021:4383418. [doi: 10.1155/2021/4383418] [Medline: 33552159]
34. Amin M, Nyachhyon P, Elyasi M, Al-Nuaimi M. Impact of an oral health education workshop on parents' oral health knowledge, attitude, and perceived behavioral control among African immigrants. J Oral Dis 2014 Jun 23;2014:1-7. [doi: 10.1155/2014/986745]

35. Sharma S, Mohanty V, Balappanavar AY, Chahar P, Rijhwani K. Role of digital media in promoting oral health: a systematic review. Cureus 2022 Sep 7;14(9):e28893. [doi: 10.7759/cureus.28893] [Medline: 36225421]
36. Goldberg E, Eberhard J, Bauman A, Smith BJ. Mass media campaigns for the promotion of oral health: a scoping review. BMC Oral Health 2022 May 14;22(1):182. [doi: 10.1186/s12903-022-02212-3] [Medline: 35568896]
37. Suresh BS, Ravishankar TL, Chaitra TR, Mohapatra AK, Gupta V. Mother's knowledge about pre-school child's oral health. J Indian Soc Pedod Prev Dent 2010;28(4):282-287. [doi: 10.4103/0970-4388.76159] [Medline: 21273717]
38. Alshammari FS, Alshammari RA, Alshammari MH, et al. Parental awareness and knowledge toward their children's oral health in the city of Dammam, Saudi Arabia. Int J Clin Pediatr Dent 2021;14(1):100-103. [doi: 10.5005/jp-journals-10005-1894] [Medline: 34326593]
39. Jumaa FA, Turki SG, Hattab KM. Mothers' knowledge toward oral health of children under 5 years old. Pakistan J Med Health Sci 2022;16(6):437-442. [doi: 10.53350/pjmhs22166437]
40. Nepaul P, Mahomed O. Influence of parents' oral health knowledge and attitudes on oral health practices of children (5-12 years) in a rural school in KwaZulu-Natal, South Africa. J Int Soc Prev Community Dent 2020 Sep 28;10(5):605-612. [doi: 10.4103/jispcd.JISPCD_273_20] [Medline: 33282770]
41. Sehrawat P, Shivlingesh KK, Gupta B, Anand R, Sharma A, Chaudhry M. Oral health knowledge, awareness and associated practices of pre-school children's mothers in Greater Noida, India. Niger Postgrad Med J 2016;23(3):152-157. [doi: 10.4103/1117-1936.190344] [Medline: 27623728]
42. Đorđević A. Parents' knowledge about the effects of oral hygiene, proper nutrition and fluoride prophylaxis on oral health in early childhood. Balkan J Dent Med 2018;22(3):26-31. [doi: 10.2478/bjdm-2018-0005]
43. Gurunathan D, Moses J, Arunachalam SK. Knowledge, attitude, and practice of mothers regarding oral hygiene of primary school children in Chennai, Tamil Nadu, India. Int J Clin Pediatr Dent 2018;11(4):338-343. [doi: 10.5005/jp-journals-10005-1535] [Medline: 30397379]
44. Das H, Janakiram C, Ramanarayanan V, et al. Effectiveness of an oral health curriculum in reducing dental caries increment and improving oral hygiene behaviour among schoolchildren of Ernakulam district in Kerala, India: study protocol for a cluster randomised trial. BMJ Open 2023 Feb 20;13(2):e069877. [doi: 10.1136/bmjopen-2022-069877] [Medline: 36806129]
45. Chawłowska E, Karasiewicz M, Lipiak A, et al. Exploring the relationships between children's oral health and parents' oral health knowledge, literacy, behaviours and adherence to recommendations: a cross-sectional survey. Int J Environ Res Public Health 2022 Sep 8;19(18):11288. [doi: 10.3390/ijerph191811288] [Medline: 36141563]
46. Rahman MM. Knowledge and practices towards oral hygiene of children aged 5-9 years old: a cross-sectional dataset among mothers visited tertiary level hospitals. figshare. 2024 Aug 31. URL: https://doi.org/10.6084/m9.figshare.26886547.v1 [accessed 2024-01-13]

Abbreviations

ECC: early childhood caries
NIPSOM: National Institute of Preventive and Social Medicine
STROBE: Strengthening the Reporting of Observational Studies in Epidemiology

Edited by E Meinert, T Leung; submitted 10.04.24; peer-reviewed by B Nwankwo, MH Islam; revised version received 13.09.24; accepted 27.11.24; published 03.02.25.

Please cite as:
Tamannur T, Das SK, Nesa A, Nahar F, Nowshin N, Binty TH, Shakil SA, Kundu SK, Siddik MAB, Rafsun SM, Habiba U, Farhana Z, Sultana H, Kamil AA, Rahman MM
Mothers' Knowledge of and Practices Toward Oral Hygiene of Children Aged 5-9 Years in Bangladesh: Cross-Sectional Study
JMIRx Med 2025;6:e59379
URL: https://xmed.jmir.org/2025/1/e59379
doi:10.2196/59379

Identifying Safeguards Disabled by Epstein-Barr Virus Infections in Genomes From Patients With Breast Cancer: Chromosomal Bioinformatics Analysis

Bernard Friedenson, PhD

Department of Biochemistry and Medical Genetics, Cancer Center, University of Illinois Chicago, 900 s Ashland, Chicago, IL, United States

Corresponding Author:
Bernard Friedenson, PhD
Department of Biochemistry and Medical Genetics, Cancer Center, University of Illinois Chicago, 900 s Ashland, Chicago, IL, United States

Related Articles:

Companion article: https://www.medrxiv.org/content/10.1101/2023.07.03.23292185v1

Companion article: https://med.jmirx.org/2025/1/e70039

Companion article: https://med.jmirx.org/2025/1/e70041

Companion article: https://med.jmirx.org/2025/1/e69307

Abstract

Background: The causes of breast cancer are poorly understood. A potential risk factor is Epstein-Barr virus (EBV), a lifelong infection nearly everyone acquires. EBV-transformed human mammary cells accelerate breast cancer when transplanted into immunosuppressed mice, but the virus can disappear as malignant cells reproduce. If this model applies to human breast cancers, then they should have genome damage characteristic of EBV infection.

Objective: This study tests the hypothesis that EBV infection predisposes one to breast cancer by causing permanent genome damage that compromises cancer safeguards.

Methods: Publicly available genome data from approximately 2100 breast cancers and 25 ovarian cancers were compared to cancers with proven associations to EBV, including 70 nasopharyngeal cancers, 90 Burkitt lymphomas, 88 diffuse large B-cell lymphomas, and 34 gastric cancers. Calculation algorithms to make these comparisons were developed.

Results: Chromosome breakpoints in breast and ovarian cancer clustered around breakpoints in EBV-associated cancers. Breakpoint distributions in breast and EBV-associated cancers on some chromosomes were not confidently distinguished ($P>.05$), but differed from controls unrelated to EBV infection. Viral breakpoint clusters occurred in high-risk, sporadic, and other breast cancer subgroups. Breakpoint clusters disrupted gene functions essential for cancer protection, which remain compromised even if EBV infection disappears. As CRISPR (clustered regularly interspaced short palindromic repeats)–like reminders of past infection during evolution, EBV genome fragments were found regularly interspaced between Piwi-interacting RNA (piRNA) genes on chromosome 6. Both breast and EBV-associated cancers had inactivated genes that guard piRNA defenses and the major histocompatibility complex (MHC) locus. Breast and EBV-associated cancer breakpoints and other variations converged around the highly polymorphic MHC. Not everyone develops cancer because MHC differences produce differing responses to EBV infection. Chromosome shattering and mutation hot spots in breast cancers preferentially occurred at incorporated viral sequences. On chromosome 17, breast cancer breakpoints that clustered around those in EBV-mediated cancers were linked to estrogen effects. Other breast cancer breaks affected sites where EBV inhibits JAK-STAT and SWI-SNF signaling pathways. A characteristic EBV-cancer gene deletion that shifts metabolism to favor tumors was also found in breast cancers. These changes push breast cancer into metastasis and then favor survival of metastatic cells.

Conclusions: EBV infection predisposes one to breast cancer and metastasis, even if the virus disappears. Identifying this pathogenic viral damage may improve screening, treatment, and prevention. Immunizing children against EBV may protect against breast, ovarian, other cancers, and potentially even chronic unexplained diseases.

(JMIRx Med 2025;6:e50712) doi:10.2196/50712

KEYWORDS

breast cancer; cancer; oncology; ovarian; virus; viral; Epstein-Barr; herpes; bioinformatics; chromosome; gene; genetic; chromosomal; DNA; genomic; BRCA; metastasis; biology

Introduction

In the United States, over 40,000 women die from breast cancer each year [1,2]. The causes of the disease are not well understood, making prevention and treatment empirical and hazardous. At the time of breast cancer diagnosis, its causes are difficult to isolate from multiple risk factors. A human cancer virus is one such risk factor. A tumor virus does not cause cancer by itself [3] but can make cancer more likely by inhibiting tumor suppressors [4] or activating oncogenes. Viral damage then increases cancer risks via mutations and chromosome breaks. Epstein-Barr virus (EBV), also called human herpesvirus 4, infects at least 90% of humans as a lifelong infection, often acquired at an early age [5], but the virus remains latent and asymptomatic in most people. EBV may be a risk factor for breast cancer. Active infection is significantly more prevalent in breast cancer tissues than in normal and benign controls [6], increasing risk by 4.75- to 6.29-fold [7]. EBV transformed human mammary epithelial cells in culture so that xenografts in immunosuppressed mice accelerated breast cancer. Once malignant transformation occurred, EBV was no longer required [8], but the cells remain malignant.

There has been no way to test the idea that EBV causes breast cancer and can then disappear. However, cancers in other tissues have proven relationships to EBV infection, so these known EBV-associated cancers can be compared to breast cancers at the genome level. Cancers with unambiguous EBV associations include nasopharyngeal cancer (NPC), EBV-positive diffuse large B-cell lymphoma (DLBCL), endemic Burkitt lymphoma (BL) [9], and gastric cancer (GC). Some genomic similarities between these EBV-associated cancers and breast cancer can be derived from the literature. In NPC, 100% of malignant cells are EBV positive [10]. Over 64% of NPCs are deficient in a pathway that depends on the breast cancer susceptibility genes *BRCA1* and *BRCA2* [11], which accurately repair DNA crosslinks and breaks via the homologous recombination pathway. This sprawling, interconnected pathway includes Fanconi anemia (FA) gene products and is often designated as the FA-BRCA pathway. In 126 patients with NPC, *BRCA1* and *BRCA2* were the most frequently mutated genes (55.5% and 33.3%, respectively) [12]. NPC mutations interfere with innate immunity and constitutively activate an inflammatory response. Overexpressed nuclear factor–κB (NF-κB) is a hallmark of NPC, occurring in 90% of NPCs [11]. Similarly, almost all stage-3 breast cancers overexpress NF-κB [13].

In NPC and the other known EBV-associated cancers, EBV inhibits the FA-BRCA pathway by various methods, including using viral microRNAs to downregulate *BRCA1* [14], hijacking other pathway components [15,16], and destabilizing SMC5/6-mediated chromatin interactions [17,18]. In GC, EBV infection and FA-BRCA pathway status are mutually exclusive [19], implying that EBV infection is approximately equivalent to disabling the FA-BRCA pathway. In DLBCL, the best prognostic marker is FA-BRCA pathway status [20]. In DLBCL and endemic BL, EBV variant infection accompanies *MYC* translocations. These translocations drive the disease and make a characteristic replacement of normal *MYC* control elements with highly active immunoglobulin regulatory sequences [21,22]. *MYC* amplification is frequent in breast cancers that have inactive *BRCA1* [23].

NPCs, DLBCLs, BLs, GCs, and breast cancers all have deficits in correctly repairing double-strand breaks and crosslinks. The compromised FA-BRCA pathway can produce chromosomes with too many centromeres. During cell division, mitotic spindles pull chromatids with multiple centromeres in too many directions, generating chromosome breaks to destabilize the human genome [24,25]. In breast cancer, these variations mark breakpoints at translocations and oncogene amplifications [26].

If EBV contributes to breast cancer, gene deficits in breast cancers and EBV-associated cancers should produce comparable changes in the human genome that do not depend on whether EBV infection persists. The aim of this study was to test for these virus-induced genome changes using bioinformatic calculations and analyses. The results could implicate EBV and its variants in disabling a variety of molecular and cellular safeguards that protect against breast cancer and its metastasis. Whether or not cancer develops in response to EBV infection depends on major histocompatibility complex (MHC) gene polymorphisms [27,28], so not everyone infected with EBV will develop cancer. In susceptible people, genome damage is permanent and does not require large numbers of viral particles, active infection, or continuing virus presence. Childhood immunization against selected EBV gene products may do much to prevent breast, ovarian, and other cancers.

Methods

Datasets Used in the Analysis

Overview

The initial data for analysis came from literature searches for studies on breast and EBV-associated cancers with large numbers of participants, unrestricted access to genome information, and complete whole-genome analysis. The first criterion for including breast cancer data was published intrachain or interchain chromosome breakpoints from high-quality, peer-reviewed publications produced by world-class laboratories. The second criterion was the availability of sufficient DNA sequence data to specify the location of these chromosome breakpoints. The third criterion was that genome sequencing had been done on samples taken before treatment began. These publicly available DNA sequence data were chosen to encompass diverse genetics, subtypes, stages, grades, morphologies, and outcomes. Initially, breast cancers were separated only broadly into those with a likely hereditary component versus those without this component. The cancers had to include typical morphologies such as ductal carcinomas, lobular carcinomas, medullary carcinomas, and

invasive carcinomas (ie, “no special type”). The included breast cancers were all primary stage-2 or stage-3 cancers. Although surgery usually removes these primary tumors, cells with only a few additional late mutations are responsible for seeding local recurrences or metastases, so primary and metastatic tumors are not very different [29]. Although the selected cancers are not a random sample representing all breast cancers [30], they are likely to have chromosome instability originating from diverse typical causes.

Specifically, the breast cancer data used came from 560 breast cancer genome sequences, familial cancer data from 78 patients, methylation data from 1538 breast cancers versus 244 controls, 243 triple-negative breast cancers, and 2658 human cancers [31-35]. Data also included 74 breast cancers from high-risk women who were typed as having *BRCA1*- or *BRCA2*-associated mutations or cancers diagnosed before the age of 40 years [36,37]. Another study of familial breast cancers contributed 65 familial breast cancers [33]. Gene breakpoints for many interchromosomal and intrachromosomal translocations and breakpoints were obtained from the COSMIC (Catalog of Somatic Mutations in Cancer) website, as curated from original publications or original articles and their supplemental informationsupplemental information [31-33]. Multimedia Appendix 1 provides a glossary of the terms used in this paper.

Breakpoints in Breast Cancers From High-Risk Women

Hereditary cancers were taken as breast cancers from women with a typed high-risk *BRCA1* or *BRCA2* mutation diagnosed before the age of 70 years. Cancers from patients with onset before the age of 50 years were also included to add more data, since these women are at high risk for an inherited, cancer-associated mutation. These patient samples were chosen based on descriptions in published data defining the breast cancer cohorts [31,33].

Sporadic Breast Cancers

Sporadic breast cancers were taken as breast cancers diagnosed after the age of 70 years that did not have a known inherited mutation [31].

Breast Cancer Subgroups

Human epidermal growth factor receptor 2 (HER2)–positive and triple-negative breast cancer data used for subgroup analysis were from original publications [33] and the COSMIC website.

Exclusions

Male breast cancers were excluded.

Data Source for Ovarian Cancers

Data for breakpoints in ovarian cancers were downloaded from the COSMIC website. The cancers corresponded to “mixed adenosquamous ovarian carcinomas” and were arbitrarily taken from those with the largest number of structural variants. These cancers all had the prefix “AOCS-” with further identification numbers and *BRCA* mutation status in parentheses as follows: 170-1-8 (negative), 120-3-6 (*BRCA2*), 142-3-5 (negative), 139-1-5 (negative), 086-3-2 (negative), 147-1-1 (*BRCA1* and *BRCA2*), 094 - 6-X (*BRCA1*), 094-1-1 (*BRCA1*), 088-3-8 (negative), 139-6-3 (*BRCA2*), 150-3-1 (negative), 116-1-3 (negative), 155-3-5 (*BRCA2*), 093-3-6 (negative), 034-3-8 (*BRCA1*), 091-3-0 (*BRCA1*), 139-19-0 (*BRCA2*), 170-3-5 (negative), 114-1-8 (negative), 064-3-3 (negative), 064-1-6 (negative), 106-1-1 (*BRCA1*), 152 - 1-X (*BRCA1*), and 134-1-5 (unknown).

Original Data Sources for Cancers With Known EBV Associations: NPCs, Lymphomas, and GCs

Overview

NPC chromosome breakpoint positions were retrieved from Bruce et al [11] for 70 primary tumors of the nasopharynx at stages 1-4C. The data came from whole-genome sequencing of “63 micro-dissected tumors, 5-patient derived xenografts, and two cell lines.” DLBCL breakpoints were collected from 88 patients with DLBCL (aged >60 y) [22]. The *MYC* breakpoints included class I and II *MYC* translocation locus breakpoints defined in BL, encompassing areas far upstream of *c-myc* [38-40]. Downstream breakpoints included an enhancer region approximately 565 kilobases long on the nearest telomere side of the *MYC* coding sequence [22]. Older data provided fusion sequences as Gencode Accession numbers [21]. These fusion sequences were downloaded as FASTA files and copied to BLAST (Basic Local Alignment Search Tool) for placement on the human GRCh38/hg38 reference sequence. GCs with inferred EBV infection status came from 34 (20.2%) out of 168 samples subjected to whole-genome sequencing [41].

Selection Bias

As much as possible, selection bias was avoided by blindly selecting samples, replicating samples using cohorts from different publications, using the largest possible groups of samples, and avoiding convenience sampling. Some experiments used a newer dataset from 780 breast cancers [22] for comparisons to confirm that selection bias was unlikely.

Recruitment

Data from genome sequence studies did not include specific recruitment procedures for patients with cancer. However, patients are typically recruited through hospitals and clinics with referrals from medical professionals. Patients provide informed consent to have their genomes sequenced and used for research and to integrate cancer genome sequence data into treatment decisions [42].

Methods Used to Determine That DNA Breakpoints From Breast and Ovarian Cancers Clustered Around Breakpoints in EBV-Associated Cancers

Calculation of Distances Between Breakpoints in Breast and Ovarian Cancers Versus EBV-Related Cancers

Before combining or comparing datasets, they were all converted to the same genome version, usually GRCh38. The break position in breast cancer nearest to a break in NPC was taken as the Microsoft Excel *XLOOKUP* value for the number of base pairs (bp) from the closest NPC breakpoint 5’ to the breast cancer break or the NPC breakpoint 3’ to the breast cancer break, whichever was closer (Multimedia Appendix 2). For comparing a given breast cancer breakpoint A2 to EBV-associated cancer breakpoints B2 to B72, the initial

algorithm to find the nearest 5' break position was written in cell C2 as follows:*=XLOOKUP($A2,$B$2:$B$72,$B$2:$B$72,0, −1,1)*. Changing −1 to +1 gave D2, the nearest 3' position. Distance from the breast cancer breakpoint was then calculated as *=MIN(ABS(C2-A2), ABS(D2-A2))*. The same formulas were then continuously updated by Excel to calculate all other breast cancer comparisons in column A. Differences in the amount of data available for NPC versus breast cancer breakpoints complicated the calculations near chromosome telomeres. Several methods of handling these end regions made no discernible difference in the outcomes. For a 5000-bp window, an overflow window of 5,000,000 was used to limit the number of bins to a maximum of 1000. Another method of calculating distances between chromosome breakpoints in different cancers used the minimum of the absolute values of distances between breast cancers and the array of breakpoints in GCs, BLs, or NPCs. This method gave results identical to *XLOOKUP* values but was more convenient to compare clusters of breast cancer breakpoints to those in lymphoid and epithelial EBV-associated cancers. Hundreds of millions of calculations were repeated at least twice. Most of the calculations in this section are presented in Multimedia Appendix 2.

DNA Sequence Homology Analyses to Determine Breakpoints in Human Cancer Sequences That Resemble Viral Sequences

The NCBI BLASTn program (MegaBLAST) and database [43-45] were used to compare DNA sequence homologies around breakpoints in breast cancers to all available viral DNA sequences. *E* ("expect") values are related to *P* values and represent the probability that a given homology bit score occurs by chance. *E* values $<1\times10^{-10}$ were considered significant homology. In many cases, *E* values were "0" ($<1\times10^{-180}$) and always far below 1×10^{-10}. The virus DNA was retrieved from BLAST searches using "viruses (taxid:10239)," with human sequences, mouse sequences, and uncharacterized sample mixtures excluded. Different strains and isolates of the same virus were tested for human homology. Specifically, the HKHD40 and HKNPC60 variants were often considered together as "EBV."

Methods Used for Chromosome Comparisons of Breakpoints in Breast Cancers in High-Risk Women Versus Breakpoints in Sporadic Breast Cancer

The NCBI Genome Decoration page provided chromosome annotation software [46].

Identifying Genes Around the Most Frequent EBV-Binding Site Locations and Tethering Sites

EBV nuclear antigen 1 (EBNA1)–binding location genome coordinates [47,48] were used to tabulate genes within or near anchoring sites where EBV docks on human DNA. Breaks in breast cancers were compared to the gene positions around their EBNA1-binding sites. The Palindrome Site Finder from NovoPro and the EMBOSS palindrome program were used to identify palindromic DNA sequences.

Comparisons for Similarities Among Human Herpesviruses

EBV variants HKHD40 and HKNPC60 were compared to human herpesviruses in BLASTn by entering the terms "human gamma herpesvirus 4," "herpesviridae," and "herpesvirales." Values with ≥2000 bp in common were selected. The EBV reference sequence was also tested against the following proven cancer viruses: human herpesvirus 8 (also called Kaposi sarcoma virus), herpes simplex virus 1, and human cytomegalovirus.

Locating Piwi-Interacting RNA Sequences as Evidence of Past EBV Infection

Piwi-interacting RNA (piRNA) locations were retrieved from the piRNA bank [49,50]. To compare the positions of piRNAs in virus homology versus genome position graphs, the midpoints of piRNA sequences were assigned arbitrary homology values. Positions of differentially methylated regions near breast cancer breakpoints on chromosome 6 [51] were compared to breakpoint positions for 70 NPCs based on published data analyses [11].

Viral Sequences in Human Genomes as Hypermutation and Rearrangement Sites in Breast Cancers

A graph of viral sequences in humans against chromothripsis breaks in breast cancers was so complex that it resisted interpretation, so only the 5 viral sequences nearest the chromothripsis breaks were used. The viral sequences nearest high-confidence chromothripsis breaks were determined in 5 iterations as genome coordinates where *XLOOKUP* values gave the minimum distances. Distances between all virus homology start points were then compared to all chromothripsis breakpoints.

Methods of Data Analyses and Statistical Software

DNA flanking sequences at breakpoints were downloaded primarily from the GRCh38/hg38 version of the University of California, Santa Cruz Genome Browser as FASTA files and copied directly into BLAST. Results were checked against breakpoints in 101 triple-negative breast cancers from a population-based study [32]. The University of California, Santa Cruz Genome Browser's *Liftover* function interconverted different versions of genome coordinates into GRCh38/hg38 coordinates.

Statistics

Excel, SPSS (IBM Corp), StatsDirect, Visual Basic (Microsoft), and Python (Python Software Foundation) scripts were used for data analysis. Mann-Whitney *U* tests compared overall breakpoint distributions [52] and tested the hypothesis that breakpoint distributions were identical or at least roughly the same. The Mann-Whitney *U* test was chosen because the comparisons involved unequal numbers of breakpoints, and each observation was likely independent. *P* values >.05 were taken to indicate that identical distributions could not be excluded. Tests for normality included kurtosis and skewness values and evaluation by Shapiro-Francia and Shapiro-Wilk methods [53] (Multimedia Appendix 2). The Fisher exact test compared breakpoints in breast cancers to those in known viral cancers. The unpaired 2-tailed Student *t* test was used to compare the means of numbers of breast cancers with severe

versus nil lymphocyte infiltrates, assuming the data approximated normality and that there were no extreme outliers. Both of these tests require independence and random sampling. All these test results are only approximate because they depend on underlying assumptions.

Fragile Site Sequence Data

Positions of common fragile sites were retrieved from a database [54] and original publications [55].

Ethical Considerations

This study presents analyses of publicly available data without recruiting additional human or animal subjects. Because this study is a secondary analysis, it is exempt from institutional review board and ethics approval. The data are in the public domain and are available for independent research and analysis [56]. It is not necessary to obtain permission to reuse public data. The original informed consent allows secondary analysis without additional permission.

Results

Breakpoints in Breast Cancers From High-Risk Backgrounds Clustered Around Breakpoints in NPC, an EBV-Mediated Cancer

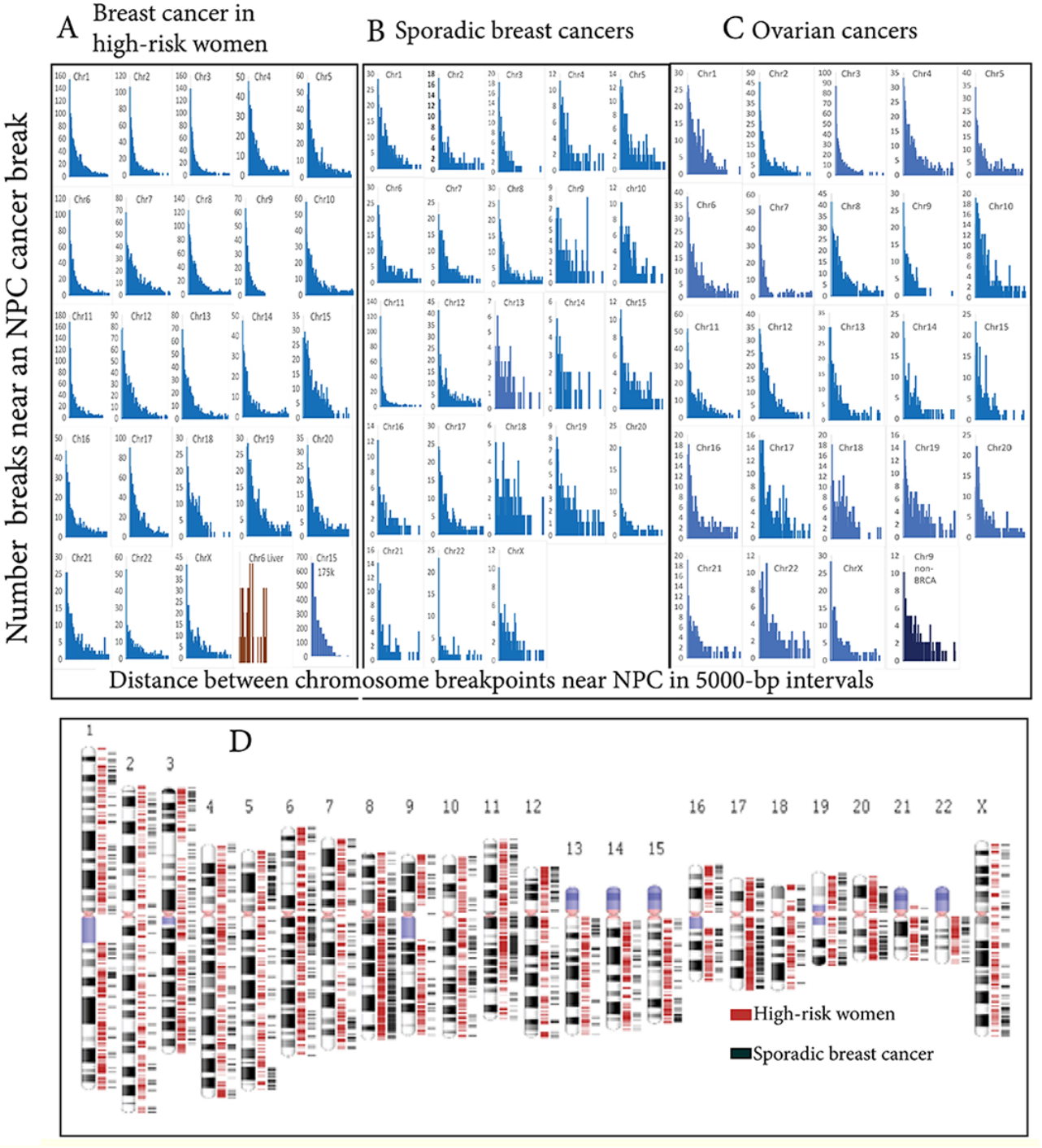
EBV-mediated cancers such as NPC have defects in DNA repair and inflammatory pathways, resembling hereditary breast and ovarian cancers. To further characterize this resemblance, breakpoints in 70 NPC genomes were compared to breakpoints in 139 breast cancer genomes from high-risk women (*BRCA1*/*BRCA2* mutation, familial concentration, or young age).

The distances from all breast cancer breakpoints to the nearest NPC breakpoints across the entire length of chromosome 1 produced results with so many points that they were difficult to interpret (Multimedia Appendix 3). Different laboratories collected these breakpoint data over many years. To allow for some variations, the data were grouped into 5000-bp increments (2×10^{-5} relative error). As shown in Figure 1 and Multimedia Appendix 2, breast cancer breakpoints were most often clustered within 5000 bp of NPC breakpoints, but many breakpoints agreed much more closely. A total of 20 breast cancer breakpoints on chromosome 1 were within 500 bp of an NPC breakpoint, and several chromosomes had breast cancer and NPC breakpoints in essentially the same positions. As represented by Mann-Whitney *U* test results (Multimedia Appendix 2), breast cancer and NPC breakpoint distributions were statistically the same ($P>.05$) for chromosomes 6, 7, 10, 13, 14, 15, 22, and X, but different on chromosome 1 and other chromosomes ($P<.05$).

In contrast, liver cancer breakpoints at hepatitis B virus integration sites [57] differed from those in breast cancer or NPC (Figure 1). No breaks in 114 liver cancers on chromosome 1 were within 5000 bp of breaks in any NPC; only one break on chromosome 6 in 61 liver cancers fit this window. According to a meta-analysis, the chance that breakpoints on chromosomes 1, 2, 6, and 8 were not within 5000 bp in liver cancer versus NPC was 4.4 (95% CI 1.9 - 10). NPC and liver cancer did not have the same breakpoint distributions ($P<.001$).

The above results revealed that breast cancer breakpoints in high-risk women were clustered near those in the EBV-associated cancer, NPC, on every chromosome. The next step was to decide whether these similarities depended on mutations in the breast cancer susceptibility genes, *BRCA1* or *BRCA2*, by comparisons to sporadic breast cancers. The sporadic breast cancer group comprised 74 women, aged ≥70 years, with normal *BRCA* genes and no other known inherited, cancer-associated mutations [31]. Like breakpoints from high-risk women, many sporadic breast cancer breakpoints clustered around those in NPC (Figure 1). Breakpoints in these sporadic breast cancers clustered at chromosomal locations similar to breast cancers from high-risk women, although the frequencies and distributions sometimes differed significantly. The patients with sporadic breast cancer were older than the high-risk women, arguing against age as responsible for similarity to NPC breakpoints.

Figure 1. (A) Breakpoints in 139 breast cancers from high-risk women (*BRCA* mutation, familial concentration, or early onset) clustered around breakpoints in 70 NPCs. The data were grouped in 5000-bp increments to allow for methodological and laboratory differences. An unrelated set of hepatocellular data associated with hepatitis B insertions did not show a similar relationship to NPC. Breast cancer and NPC breakpoint distributions could not be confidently distinguished (P>.05) for chromosomes 6, 7, 10, 13, 14, 15, 22, and X (Multimedia Appendix 2). Many breakpoints were virtually the same on some chromosomes. The panel at the lower right shows how the selection of a larger bin size of 175,000 bp (the approximate length of EBV) affects the distributions of breakpoints. (B) Like the breast cancers from high-risk women, breakpoints in 74 sporadic breast cancers clustered around the breakpoints found in 70 NPCs. Breast cancer breakpoints within 5000 bp of an NPC breakpoint were the largest single category on most chromosomes. (C) Breakpoints in 25 mixed adenosquamous ovarian cancers also clustered around breakpoints in the 70 NPCs. The data show both *BRCA*-associated and nonassociated ovarian cancers. The panel in the lower right corner represents chromosome-9 data after removing all *BRCA*-associated ovarian cancers. The sporadic cancers show the same results as the complete set but with less data. (D) Many breakpoints in sporadic breast cancers clustered at chromosomal locations similar to those from high-risk women. Interchromosome translocation break positions in 74 mutation-associated, familial, or early-onset female breast cancers (red) versus 74 likely sporadic female breast cancers (black) are shown. bp: base pairs; Chr: chromosome; EBV: Epstein-Barr virus; NPC: nasopharyngeal cancer.



Viral Homologies Around Breakpoints in Mixed Adenosquamous Ovarian Carcinoma Also Clustered Around Breakpoints in EBV-Mediated Cancer

Ovarian cancer data enabled an additional test for EBV involvement in breast cancer because, like breast cancer, *BRCA1* or *BRCA2* mutations can also predispose patients to ovarian cancer [58]. Chromosome breakpoints in 25 mixed adenosquamous ovarian cancers were compared to breakpoints in NPCs. The results depicted in Figure 1 emulated breast cancer comparisons. Nearly half (12/25, 48%) the ovarian cancer cases had likely hereditary *BRCA* mutations. The remaining sporadic ovarian cancers gave the same results as the complete set but with less data. As in breast cancer, ovarian cancer breakpoint distributions clustered around NPC breakpoints, even without a hereditary *BRCA1* or *BRCA2* gene mutation driver.

Breaks in Lymphomas Associated With EBV Infection Also Matched Breast Cancer and NPC

EBV drives lymphomas as well as NPCs. Based on epidemiologic research results, FA-BRCA pathways protect against lymphomas [59,60]. If EBV is genuinely associated with breast cancer breakpoints, then breakpoint positions in EBV-mediated lymphomas should also resemble those of breast and ovarian cancers. Because *MYC* gene rearrangements are characteristic of EBV-associated lymphomas, the first test of this idea was to survey virus-like sequences surrounding the *MYC* gene locus on the human reference genome. Figure 2 shows that *MYC* resides in a literal forest of retrovirus sequences (eg, human immunodeficiency virus type 1 [HIV1], feline leukemia virus, porcine endogenous retrovirus, and human endogenous retrovirus [HERV]) interspersed with EBV-like sequences.

The concentration of virus sequences around *MYC* on chromosome 8 prompted the addition of the EBV-associated lymphoma DLBCL to breakpoint comparisons. As shown in Figure 2, the results revealed that hundreds of breast cancer and NPC breakpoints congregated around breakpoint positions in 88 DLBCLs [22]. This agreement was consistent with other similarities between breast cancers and these EBV-associated cancers, including deficits in FA-BRCA pathway–mediated DNA repair by homologous recombination [61] and NF-κB activation [11,62-64].

EBV is also a proven driver of at least one subset of BLs, typically those with *MYC* translocations. BL subsets can have mutations that impair homologous recombination [65], so results in Figure 2 revealed many breast cancer breakpoint positions near corresponding BL breakpoints. An older dataset from BLs [21] had translocation breakpoints in the virus sequence–rich area near the *MYC* locus, agreeing with about 140 breast cancer breakpoints. Four different NPC breakpoints produced over 100 matches to BL translocation breakpoints, beginning at 8250 bp apart. An unpaired, 2-tailed *t* test did not support a statistically significant difference between BL and NPC breakpoints in this area (P=.69).

Further tests were conducted to determine whether the functions of genes near clustered breakpoints supported a relationship between breast cancers and EBV-related cancers (GC [41], BL, and NPC). As illustrated in Figure 3, breast cancer breakpoints on chromosomes 6, 8, 11, and 17 aggregated near positions where breakpoints occurred in EBV-associated cancers. Many aggregated breakpoints were in the same areas as genes that control inflammation, antiviral defenses, apoptosis, intermediate filaments, epigenetic and chromatin regulation, estrogen receptor activity, mitotic structures, and mitotic controls (Table S1 in Multimedia Appendix 4). Breast cancer breakpoints that clustered around EBV-associated cancer breakpoints were especially numerous on chromosome 17. One of these clusters marked in Figure 3 included the HER2 amplicon and the topoisomerase 2a gene, with *BRCA1* and *SMARCE1* genes nearby. *SMARCE1* encodes a part of a chromatin regulation complex. Chromosome 17 breakpoints near *CNTROB* and *CTC1* genes connect EBV to centriole and telomere malfunctions during mitosis (Table S1 in Multimedia Appendix 4). Rearrangements near breakpoints may cause over- or underexpression of nearby genes (Table S2 in Multimedia Appendix 4). Many additional correlations were also likely revealed in Figure 3 but were not investigated further.

Results in this section show that breast cancer breakpoints clustered around breakpoints in additional EBV-associated cancers, where they affect critical functions needed to prevent breast cancer. Once these functions are compromised, cancer can occur without the continuing presence of EBV.

Figure 2. (A) Human DNA around the *MYC* locus on chromosome 8 was filled with virus-like sequences. *CASC11* is an RNA gene that several cancers overexpress. Breast cancer and lymphoma breakpoints were dispersed throughout the *MYC* region and beyond, but NPC breakpoints were less common. (B) On chromosome 8, hundreds of breakpoints in breast cancers and NPCs clustered around breakpoints in data from 88 patients with DLBCL who were likely EBV positive. This agreement highlights multiple similarities among these cancers. (C) EBV drives a subset of BLs, typically with *MYC* translocations and impaired homologous recombination. Based on *MYC* fusion sequences in BL, breast cancer breakpoints on chromosome 8 also clustered around BL breakpoints. BLs from an older dataset [21] had translocation breakpoints in the virus-rich area near the *MYC* locus, agreeing with ≥140 breast cancer breakpoints. *MYC* locus translocations had not been reported in NPCs, but NPC breakpoints still clustered around BL fusion breakpoints, although at greater distances. Four different NPC breakpoints produced over 100 matches to BL translocation breakpoints beginning at about 8250 bp apart. An unpaired, 2-tailed *t* test did not support a statistically significant difference between BL and NPC breakpoints in this area (*P*=.69). BL: Burkitt lymphoma; bp: base pairs; DLBCL: diffuse large B-cell lymphoma; EBV: Epstein-Barr virus; FeLV: feline leukemia virus; HERV: human endogenous retrovirus; HERVK: human endogenous retrovirus K; HIV1: human immunodeficiency virus type 1; HPV18: human papillomavirus 18; HRV: human retrovirus; NPC: nasopharyngeal cancer; PERV: porcine endogenous retrovirus; Stealth: stealth virus 1.

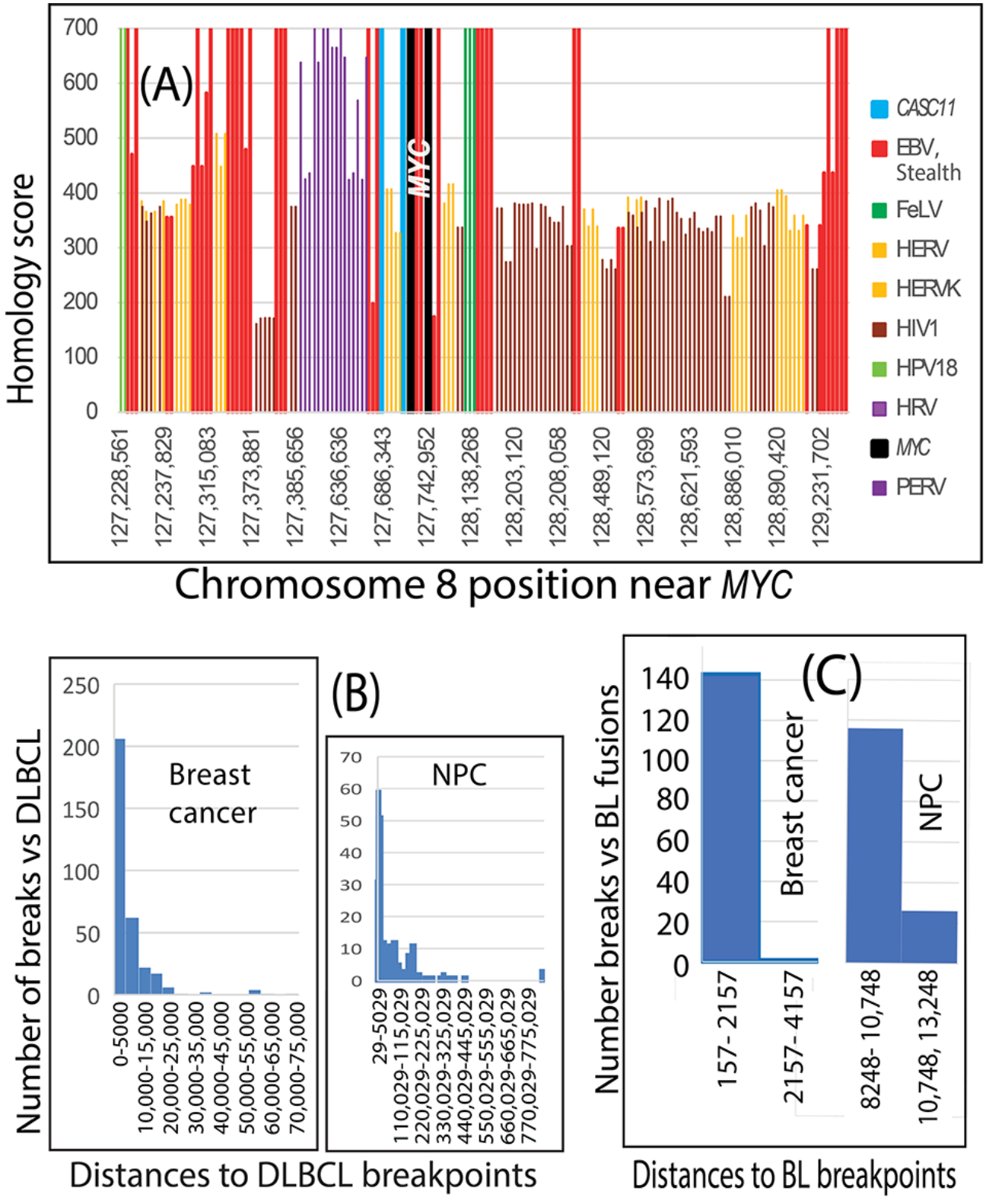
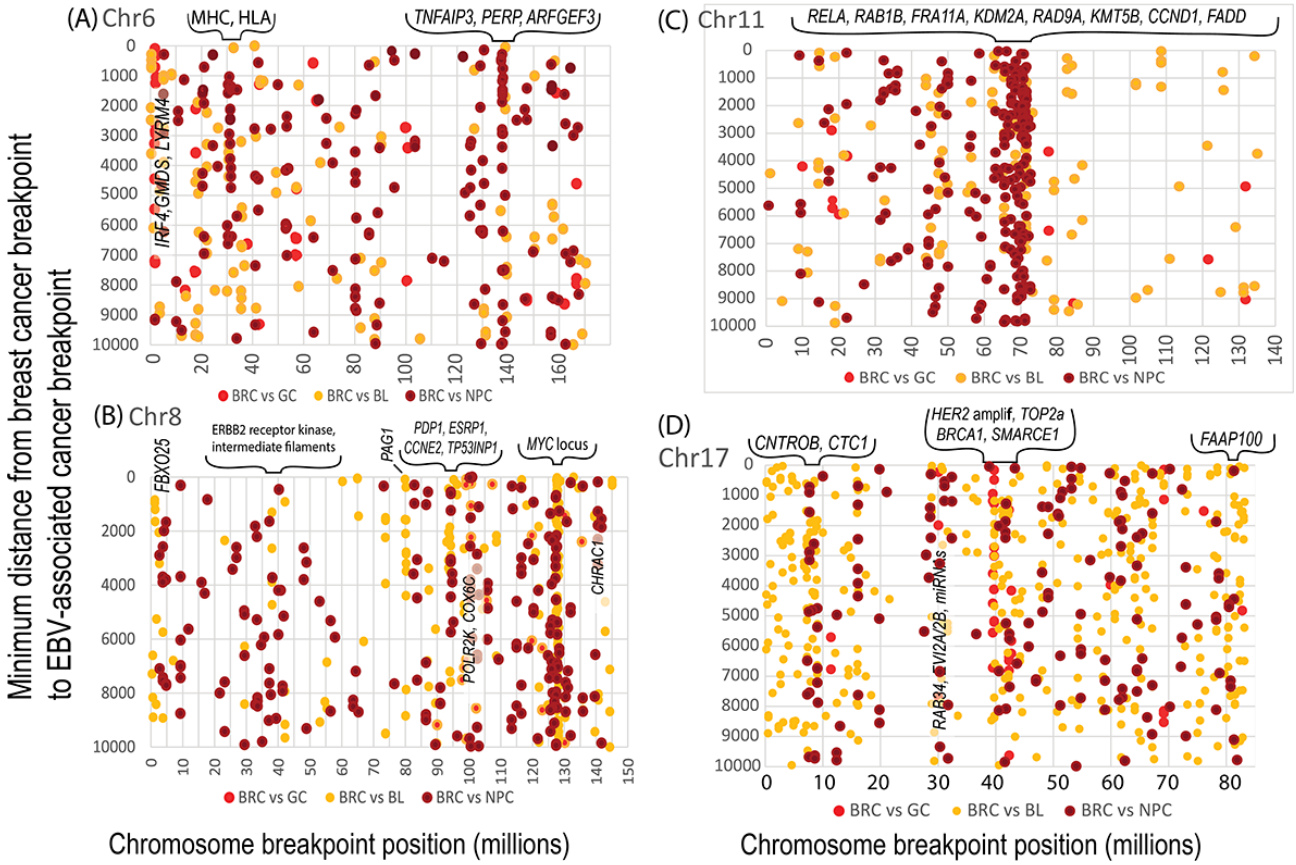


Figure 3. Breakpoints in breast cancers clustered around breakpoints in EBV-positive cancers in 3 different tissues. The EBV-positive cancers comprised 34 GCs, 90 BLs, and 70 NPCs. The clustering of breast cancer breakpoints and EBV-related cancer breakpoints was pronounced on chromosomes (A) 6, (B) 8, (C) 11, and (D) 17. Selected genes around some of the clustered breaks are indicated. Functions of the genes can have profound effects on the human genome and are summarized in Table S1 in Multimedia Appendix 4. BL: Burkitt lymphoma; BRC: breast cancer; Chr: chromosome; EBV: Epstein-Barr virus; GC: gastric cancer; HLA: human leukocyte antigen; MHC: major histocompatibility complex; NPC: nasopharyngeal cancer.



Genes at the Most Frequent EBV-Tethering Sites Clustered Around Breast Cancer Breakpoints

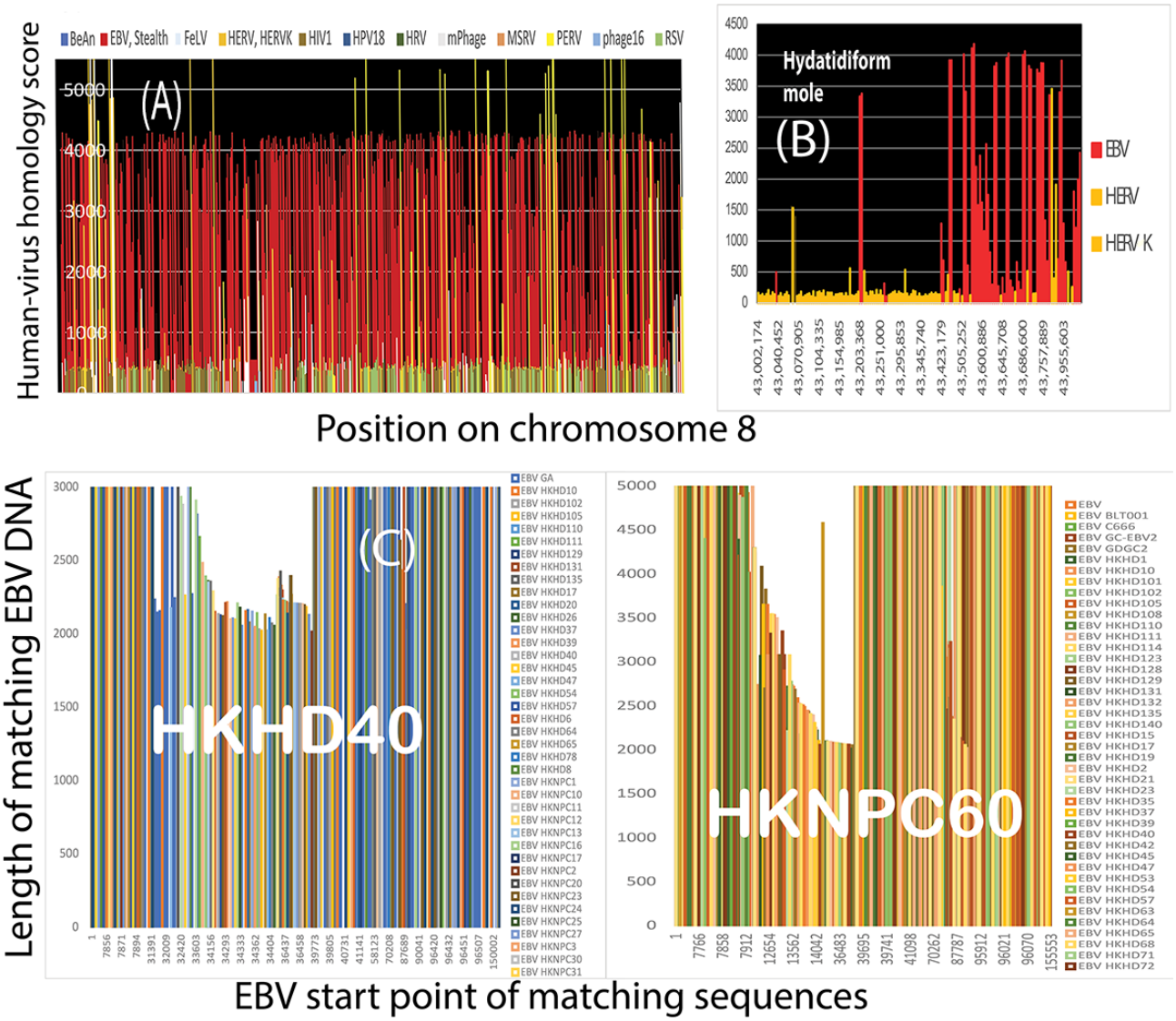
In preceding sections, breast and ovarian cancer breakpoints were found to distribute most frequently near characteristic sets of breakpoints associated with EBV-related cancers. The virus first attaches its EBNA1 protein to human DNA in the nucleus. Then, circular EBV episomes dock to this attached EBNA1 anchor. To test whether the initial EBNA1 attachment sites were related to breast cancer chromosome breakpoints, breast cancer breakpoints were compared to genes near EBV-docking sites. EBV-positive BL cells providing the data had up to 1569 EBV-docking sites on all chromosomes identified by 4C-chromatin capture experiments [47]. As shown in Figure 4A, the largest numbers of breast cancer breakpoints on most chromosomes clustered around the genes [47] nearest to genes at EBV-docking sites. In support of these comparisons, graphical estimation of virus-tethering sites on chromosome 2 from chromatin capture data for these EBV-positive cells also agreed with breast cancer breakpoints (Figure 4A). In an unrelated study [48], EBV-docking sites on chromosome 11 near known EBV anchor sites at the *FAM-D* and *FAM-B* genes were found near groups of breast cancer breakpoints, but imperfect palindrome sequences [66] were more distant (Figure 4B). This finding independently supports the idea that EBV-docking sites are near breast cancer breakpoints. Results in this section raise the possibility that EBV directly contributes to breast cancer chromosome breakpoints and fragmentation.

Figure 4. Relationships of EBV-docking sites to breast cancer breakpoints. (A) Breast cancer breakpoints clustered around the top 10% most frequently found genes near EBV-tethering sites in BL cells. Some of the best information on EBV-docking sites comes from 4C-chromatin capture experiments in EBV-positive BL cells [47]. The largest number of breast cancer breakpoints on most chromosomes clustered around the genes nearest EBV-tethering sites. BL cells providing the data had up to 1569 EBV-docking sites distributed over all chromosomes [47]. EBV-docking sites on chromosome 11 near the *LUZP2* and *FAT3* genes in BL cells were millions of bp from the 18-bp imperfect palindrome interval. Graphical estimation of virus-tethering sites on chromosome 2 (green) from these EBV-positive cells also agreed with breast cancer breakpoints. (B) Independent evidence relating breast cancer chromosome breakpoints to EBV-docking sites. Maximum homology to human DNA for all viruses (y-axis) is plotted around known EBV genome anchor sites on chromosome 11 near the *FAM55D* and *FAM55B* gene coordinates. A posited imperfect palindrome sequence [66] as an EBV-docking site was more distant from the *FAM55* genes. BL: Burkitt lymphoma; bp: base pairs; Chr: chromosome; chrom: chromatin; EBV: Epstein-Barr virus; HERV: human endogenous retrovirus; HERVK; human endogenous retrovirus K; HIV1: human immunodeficiency virus type 1; HTLV1: human T-cell lymphotropic virus type 1; MSRV: multiple sclerosis retrovirus; RSV: respiratory syncytial virus.

Breakpoints Occurred Near Human Sequences That Resemble Viruses in All Breast Cancers Tested

To further test whether EBV itself has some role in breaking chromosomes or altering their structures, human chromosomes were compared to all known viruses. As shown in Figure 5A, the results showed that nearly every breast cancer likely had undergone breakages near EBV-like sequences. Chromosome 8 alone had 59,566 significant (>200) viral homology scores. Based on data from 128 patients with breast cancers and 43,491 unique breakpoints, breakpoints in 123 (96.1%) out of 128 breast cancers were within 10,000 bp of a virus sequence. In 106 patients, the virus was an EBV tumor variant (HKHD40 or HKNPC60) with 3086 matching human sequences. According to the Fisher exact test, chromosome 8 breakpoints and EBV variant sequence matches were not independent (P<.001).

Figure 5. (A) All viral homologies on the entire lengths of chromosome 8 (a total of 145,138,636 bp) are shown in 200k-bp increments. Maximum homology scores over 4000 for human DNA versus herpes viral DNA were abundant. The 4000 score corresponds to 97% human-virus identity over nearly 2500 bp, with *E* ("expect") values (essentially *P* values) effectively equal to 0. The EBV tumor variants, HKNPC60 and HKHD40, were nearly identical to human breast cancer DNA at many positions throughout chromosome 8. (B) It is unlikely that homologies to EBV sequences occurred because the human reference genome was contaminated with EBV episomes. Homozygous hydatidiform mole cells that had lost the paternal chromosomes after fertilization still had strong homology to EBV sequences, such as HKHD40 and HKNPC60 variants. (C) EBV variants HKHD40 and HKNPC60 are typical of hundreds of other EBV variants. Hundreds of human gamma herpesvirus 4 variants are almost identical to HKHD40 and HKNPC60 over at least 2000 bp. The matching sets of viruses included many high-risk herpesvirus isolates from NPCs [67]. BeAn: BeAn 58058 virus; bp: base pairs; EBV: Epstein-Barr virus; FeLV: feline leukemia virus; HERV: human endogenous retrovirus; HERVK: human endogenous retrovirus K; HIV1: human immunodeficiency virus type 1; HPV18: human papillomavirus 18; HRV: human retrovirus; mPhage: mycolicibacterium phage J1; MSRV: multiple sclerosis retrovirus; NPC: nasopharyngeal cancer; PERV: porcine endogenous retrovirus; RSV: respiratory syncytial virus; Stealth: stealth virus 1.



Many areas on other chromosomes also had 97% human-virus identity over nearly 2500 bp. It is implausible that this much similarity comes from EBV DNA being carried over into the human reference genome. Viral homology occurred with only a small, select portion of viral DNA [68]. Viral homologies were determined for a human genome in a homozygous karyotype, haploid cell line (46,XX) hydatidiform mole derived only from the paternal chromosomes in an X-bearing sperm cell after fertilization [69]. Results still showed extensive homology between the mole and EBV variants HKHD40 and HKNPC60 (Figure 5B).

HKHD40 and HKNPC60 variant sequences kept appearing in comparisons to human sequences, so these variants were tested against other herpesviruses to determine whether they were unusual. Hundreds of human gamma herpesvirus 4 variants were almost identical to HKHD40 and HKNPC60 over at least 2000 bp (Figure 5C). The matching sets of viruses included many high-risk herpesvirus isolates from NPCs [67]. Based on this information, HKHD40 and HKNPC60 strongly resembled other herpesvirus isolates, including many that confer high risks for NPC [10]. These results show that humans have interacted extensively with EBV; the results are not due to EBV impurities in the human reference genome, and the human genome has had close relationships with oncogenic EBV forms.

Evidence of Past EBV Infection

The evidence thus far supports a central hypothesis that EBV disables tumor suppressor mechanisms in breast cancer and can then disappear. This absence of viral particles is a significant experimental obstacle to testing this hypothesis. Unlike retroviruses, EBV and its variants do not have integrase enzymes, so EBV has no conventional way to insert itself into the human genome. EBV rarely integrates, with only one or two copies in BL cell lines [70].

BLAST analysis found about 65,000 areas of strong homology ($E<1\times10^{-10}$) between the human reference genome and EBV. Because 65,000 is far more than realistic EBV integration events, it suggested the possibility that some EBV sequences were fragments created by a human version of the bacterial CRISPR (clustered regularly interspaced short palindromic repeats) system. As shown previously in Figure 3, breast cancers have breakpoints that cluster around breakpoints in EBV-associated cancers and involve MHC genes.

MHC genes are encoded on chromosome 6p21.3 in a region that becomes a candidate for such a human CRISPR version. Variants of human leukocyte antigens (HLAs) in the MHC are strong risk factors for NPC infections [71] because HLAs are required to break down and display fragments of some antigens to the immune system. A total of 13 breast cancers listed on the COSMIC website had a deletion near this HLA region. About 23% of breast cancers had mutations directly affecting HLA class I or II genes. Many more breast cancers had indirect connections because they had damage to multiple genes that interact with HLAs or were otherwise essential for immunity. The MHC region also holds *NFKB1L1*, a negative regulator of the NPC overexpressed gene hallmark, NF-κB. The 139 breast cancers from high-risk women had 284 breakpoints at chromosome 6p21.3. Breakpoints in the 70 NPC cancers also clustered there, with 40 breakpoints within the 27,865,296-34,017,013 segment on chromosome 6. Variability in the inactivation of MHC genes reflects the extreme diversity of this region.

In general, the bacterial CRISPR/Cas system loosely resembles the human piRNA system, so the distribution of piRNAs was graphed. As shown in Figure 6A, hundreds of piRNA sequences cluster near the MHC region (at ~29.7 - 33.3 megabases). The piRNA system is known to inactivate virus-derived transposons (related to HERVs) by methylating or cleaving them. The distribution of piRNA fragments was then compared to the distribution of viral DNA fragments in the MHC region of chromosome 6. Figure 6B-F reveals striking similarities in how remnants of exogenous and endogenous viruses distribute relative to piRNAs. Remnants of both virus types were homologous to the same human sequence, and both types were interspaced between piRNA sequences, sometimes right next to each other. Most of these sandwiches were at a regular interval or a multiple of a regular interval.

This interspaced arrangement looked so much like CRISPR that it raised the question of whether piRNA defense mechanisms have inactivated some EBV variants in addition to their canonical role with endogenous viruses. Long stretches of endogenous transposon-like DNA sequences routinely matched exogenous viruses. As shown in Figure 6C and E, the same human DNA interval had homology both to endogenous transposons (HERV) and exogenous viral sequences (EBV variants, stealth virus 1, chikungunya virus, BeAn 58058 virus, human papillomavirus [HPV] 16, HIV1, and HERV). This result shows that the piRNA system can store the same piece of DNA to protect DNA against these different viruses.

Chromosome 6p21.3 also contains an EBV infection marker [72]. The marker was examined in 1538 breast cancers using existing methylation data [34]. As indicated in Figure 7, promoter methylation differed significantly from normal controls in the segment shown (30,523,984 - 33,216,811 on chromosome 6). Hypermethylation occurred on *STK19*, a MHC class III gene for RNA surveillance [73,74]. Hypermethylation also occurred on a gene for preventing tumors (*TNFB*) [75] and a gene for responding to antigen-antibody complexes (*C2*). Polymorphisms in *HLA-DMB* antigen and *SAPCD1*, another class III MHC gene [76], at chromosome 6p21.3 had links to Kaposi sarcoma [77]. Human herpesvirus 8 (Kaposi sarcoma virus) is a Kaposi sarcoma driver and is closely related to EBV.

These results reveal that EBV has been attacking human DNA during evolution. There is a piRNA defense mechanism for human DNA near critical immune system genes, but both EBV-associated cancers and breast cancers inactivate some of the genes that guard piRNA defenses. The histocompatibility antigen gene region of chromosome 6 can be extensively fragmented in EBV-associated and breast cancers. MHC genes have the largest number of polymorphic forms in the human genome. This variation creates differences in viral susceptibility and inactivation. Even though most people are infected, not everyone will get an EBV-related disease or cancer.

Figure 6. The human genome organizes piRNA sequences into clusters near the MHC region of chromosome 6 (6p21.3 at ~29.7-33.3 megabases), with hundreds of piRNAs nearby. (A) The levels of various piRNAs varied by more than 1000-fold, but the most abundant piRNAs were the only ones present in every cell. These abundant sequences drive the inactivation of foreign DNA. Rare piRNAs do not function in every cell but can potentially adapt to new genome invaders. (B-F) Arbitrarily selected areas of the chromosome region where piRNAs are most abundant. piRNAs were assigned sufficient homology scores to mark their positions relative to positions with homology to viruses. (C and E) Remnants of both exogenous and endogenous virus types were homologous to the same human sequence, and both types were sandwiched between piRNA sequences, sometimes right next to each other. Most sandwiches were at a regular interval or a multiple of a regular interval. The same human DNA interval has homology to endogenous transposons (HERV) and exogenous viral sequences (ChikV, HIV1, Stealth, BeAn, and HPV16). The piRNA system can store the same piece of DNA to protect DNA against these different viruses. BeAn: BeAn 58058 virus; ChikV: chikungunya virus; Chr: chromosome; EBV: Epstein-Barr virus; HERV: human endogenous retrovirus; HERVK; human endogenous retrovirus K; HIV1: human immunodeficiency virus type 1; HPV16: human papillomavirus 16; MHC: major histocompatibility complex; PERV: porcine endogenous retrovirus; piRNA: Piwi-interacting RNA; Stealth: stealth virus 1.

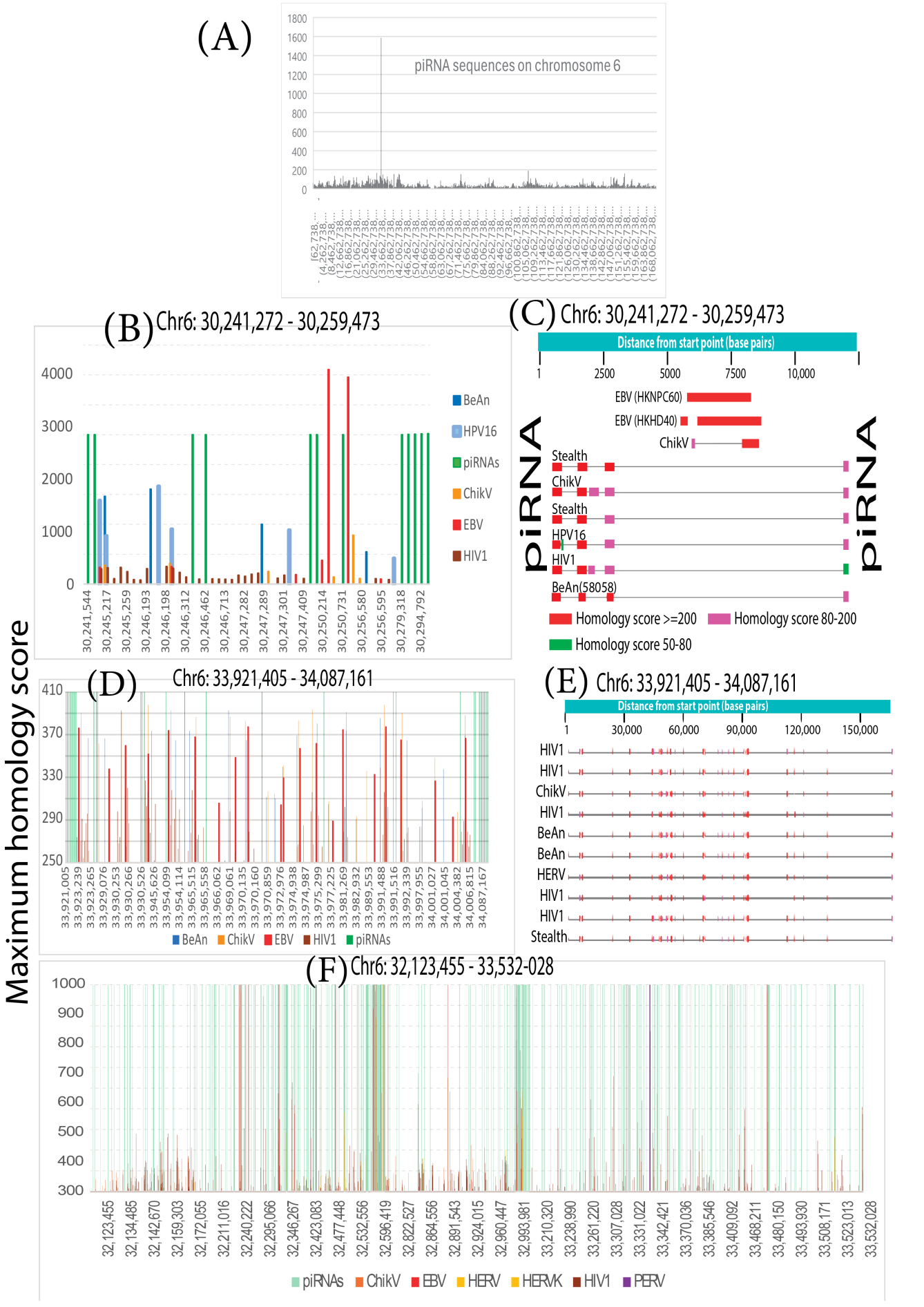
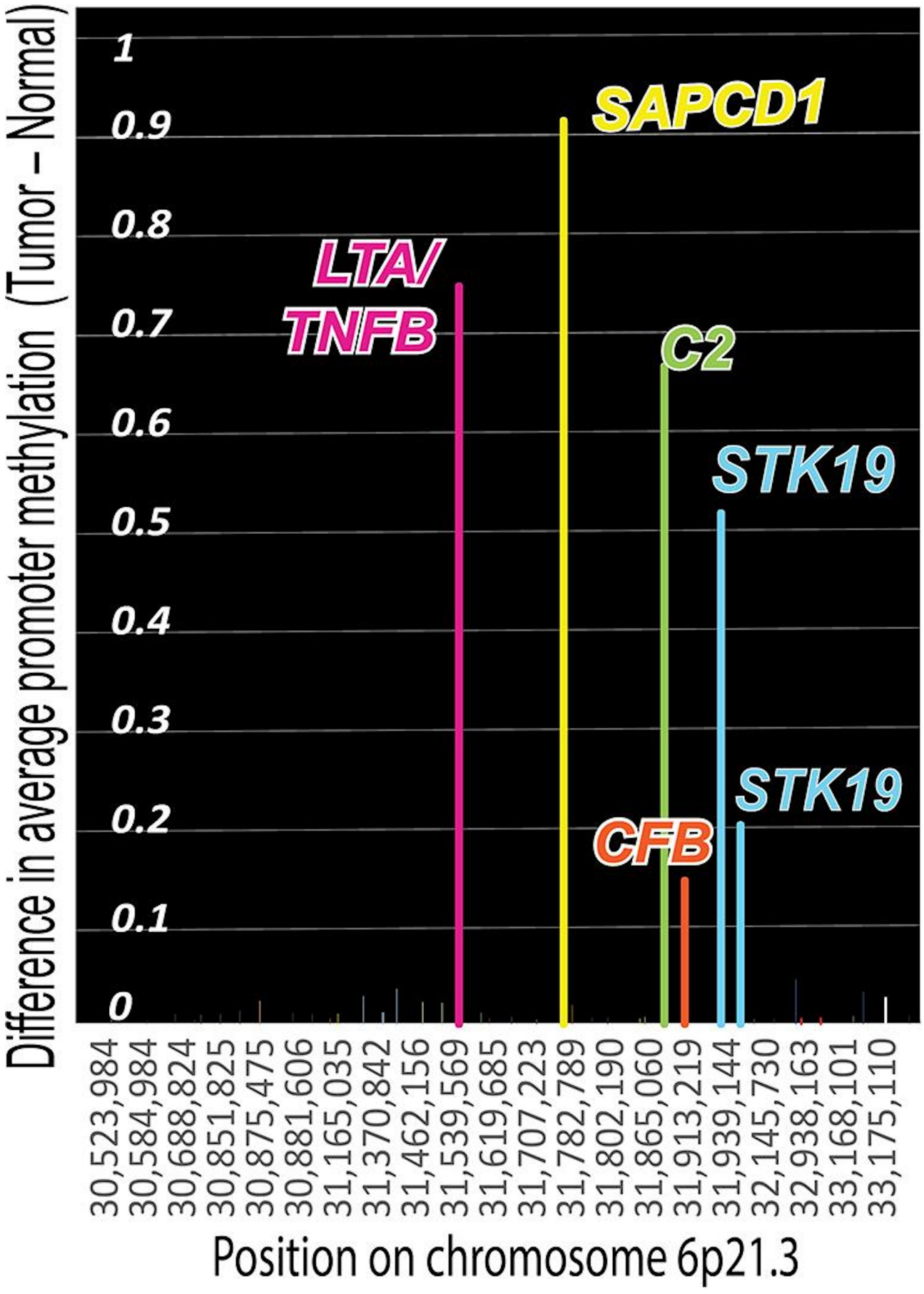


Figure 7. Chromosome 6p21.3 contains an EBV infection signature [72]. Using existing methylation data [34], the marker was examined in 1538 breast cancers. Promoter methylation in this marker region differed significantly from normal controls. Hypermethylation occurs on *STK19*, an MHC class III region gene [73] for RNA surveillance [74]. Hypermethylation also inhibited *LTA/TNFB*, a gene for preventing tumors [75], and *C2*, which encodes antigen-antibody complex responses. Polymorphisms in *HLA-DMB* antigen and *SAPCD1*, another class III MHC gene [76], at chromosome 6p21.3 have links to Kaposi sarcoma [77]. HHV8 is a Kaposi sarcoma virus closely related to EBV. EBV: Epstein-Barr virus; HHV8: human herpesvirus 8; HLA: human leukocyte antigen; MHC: major histocompatibility complex.

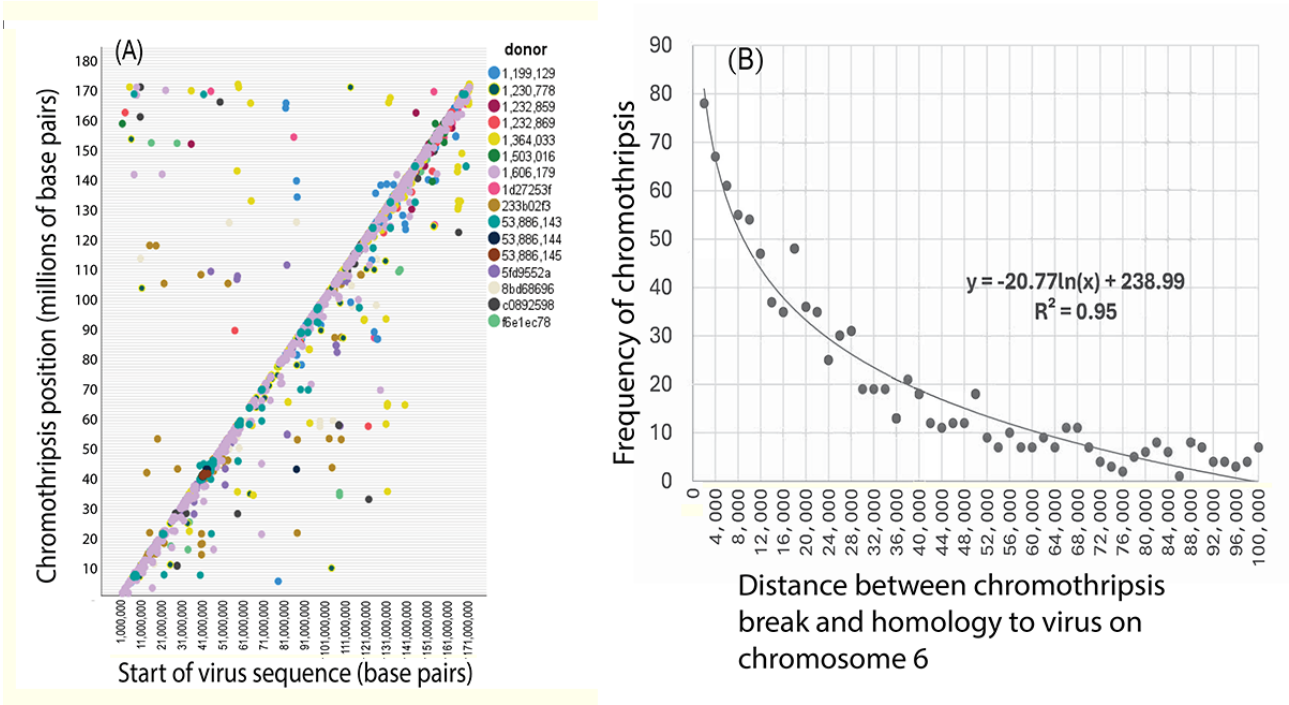


Viral Sequences in Human Genomes as Hypermutation and Rearrangement Sites in Breast Cancers

The next question was whether EBV or other virus-like sequences in the human genome cause multiple rearrangements and clustered hypermutations (chromothripsis). As shown in Figure 8A, many positions on chromosome 6 where chromothripsis occurs [35] congregated around virus sequence start positions. A total of 1090 genome coordinates described chromothripsis fragments with copy number ≥3. These coordinates were unlikely to be random since they did not follow a normal distribution (P<.001). By simple linear regression analysis (R^2=0.93), many viral sequence coordinates strongly correlated with chromothripsis positions. Figure 8B shows that as you move further away from a chromothripsis breakpoint, the frequency of breast cancer homology (score>500) to viruses decreases. This result implicates viral sequences as preferred sites where breast cancer chromosomes begin to fall apart. The equation shown mathematically describes the relationship between chromothripsis frequency and distance from viral sequences, and the constant in the equation suggests a baseline level of breakpoints.

These results suggest that homologous virus sequences at multiple positions could confuse DNA repairs already compromised by EBV in breast cancer and contribute to chromothripsis and clustered rearrangements.

Figure 8. Repetitive copies of virus sequences may confuse compromised DNA repairs and contribute to hypermutation clusters and rearrangements. (A) High-confidence positions where chromosome 6 shatters in 16 breast cancer genomes [35] were plotted against start points of viral sequence homologies. EBV or other viruses then cause groups of rearrangements and hypermutation clusters (chromothripsis). A total of 1090 genome coordinates described fragments with copy number ≥3. These coordinates were unlikely to be completely random since they did not follow a normal distribution (P<.001). Genome coordinates on chromosome 6 matching virus sequences were strongly correlated by simple linear regression analysis (R^2=0.93). (B) As you move further from chromothripsis breakpoints, the frequency of breast cancer homology to viruses decreases, according to the equation shown. The constant in the equation suggests a baseline level of breaks.



EBV and Metastasis

The last question was whether EBV contributes to breast cancer metastasis. According to Yates et al [29], relapsed and metastatic breast cancer tumors keep their tumor-driver gene mutations and continue acquiring new ones. Late mutations in JAK-STAT and SWI-SNF signaling pathways drive established breast cancers into metastasis.

NPC often loses type-1 interferon genes (*IFNA1*, *IFNA2*, *IFNA8*, and *IFNE*) and nearby *MTAP* (32%-34% [11]) by homozygous deletions at chromosome 9p21.3. Interferons initiate canonical JAK-STAT signaling by binding to cell surface receptors that then activate internal Janus kinases (JAKs). The activated JAKs phosphorylate cytoplasmic STAT (signal transducer and activator of transcription) proteins, which travel to the cell nucleus to activate interferon-responsive genes. The percentages of breast cancers on the COSMIC website with mutations in a "JAK" or "STAT" isoform or transcript variant were calculated: 7.8% had a JAK mutation and 36.7% had a STAT mutation. Deletions of interferon genes in NPC also facilitate viral replication and block interferon from activating JAK-STAT signaling. Breast cancers (Multimedia Appendix 2) have 65 breakpoints strictly within this interferon-*MTAP* region (21,579,478 - 20,503,534 on chromosome 9), not counting longer fragments that include the interval. As shown in Figure 9, breast cancer breakpoints align well with EBV-associated cancer breakpoints near the large cluster of interferon genes on chromosome 9.

Figure 9. Damage to JAK-STAT and SWI-SNF signals pushes breast cancer into metastasis [29]. EBV interferes with these signaling pathways to facilitate viral replication. (A) Breakpoints in breast cancers on chromosome 9 facilitated viral replication and blocked sources of JAK-STAT signaling, including a large cluster of interferon genes on chromosome 9. Breast cancers can disable SWI-SNF by targeting *ARID* genes. (B) *ARID1A* was encoded on chromosome 1 near a hot spot where multiple breast cancer breakpoints approximately aligned with breakage points in EBV-associated cancers. Another site at about 150,000,000 bp had a histone-rich region nearby. SWI-SNF affects histones, which also profoundly affects metastasis [78]. The GRCh38 genome version does not include centromere sequences due to technical limitations. *ANXA1*: Annexin A1; BL: Burkitt lymphoma; bp: base pairs; BRC: breast cancer; Chr: chromosome; EBV: Epstein-Barr virus; GC: gastric cancer; NPC: nasopharyngeal cancer; SWI-SNF: switch/sucrose non-fermentable.

Mutations in EBV-associated cancers show that Yates metastasis driver gene damage accompanies EBV infection. SWI-SNF (switch/sucrose non-fermentable) is a complex that repositions nucleosomes and supports genome stability [79]. SWI-SNF addresses obstacles to replication sensed by the FA-BRCA pathway [79,80]. Referring back to Figure 3, clustered breast cancer breakpoints on chromosome 17 around EBV breakpoints affect the SWI-SNF component *SMARCE1*. In addition, breast cancers can disable SWI-SNF by targeting *ARID* genes [29]. *ARID1A* is a COSMIC top-20 most frequently mutated gene in breast cancer. Like breast cancer, NPC has multiple recurrent aberrations in *ARID1A* genes. As shown in Figure 9, *ARID1A* lies near a hot spot where multiple breast cancer breakpoints approximately aligned with breakage points in EBV-associated cancers. The loss of *ARID1A* activates Annexin A1, which aligned closely with a region targeted by EBV-associated cancers on chromosome 9. A chromosome-1 site at about 150 million bp had a nearby histone-rich gene region. Histones are chromatin structures that SWI-SNF dynamically remodels to regulate access to genetic information. Histones can profoundly affect metastasis [78]. Figure 9 also reveals many additional alignments between breakpoints in breast and EBV-associated cancers that were not investigated further.

NPC often inactivates SWI-SNF components *BAP1* and *PBRM1* within a frequently damaged 3p21.3 gene cluster [11] at 52,400,000 - 53,000,000 on chromosome 3. Analyses of breast cancers found 18 breakpoints within this short interval. DLBCL, another EBV-linked cancer, also had recurrent alterations in components of SWI-SNF complexes [81].

The Warburg effect (oxidative glycolysis) [68] favors metastasis. The Warburg effect occurs in NPC because pyruvate dehydrogenase *(PDHB)* genes on chromosome 3p are deleted or rearranged in almost all cases. Similar changes to chromosome 3p were found in breast cancers, which also undergo the Warburg effect [68]. This Warburg metabolic switch favors metastasis because it mitigates oxidative stress on cancer cells. Large amounts of lactate accumulate in the absence of *PDHB* to acidify the tumor microenvironment and interfere with the destruction of metastatic cells [82].

This section's results show that EBV may push breast cancer into metastasis by interfering with JAK-STAT and SWI-SNF signaling pathways to facilitate viral replication while making the microenvironment more favorable to tumor growth.

Alternative Explanations for Breast Cancer Breakpoints That Do Not Involve EBV Variants

Subgroups

To determine whether breakpoint similarities in viral and breast cancers depended on specific subgroups, relationships to NPC were compared in triple-negative and HER2-positive breast cancers (20 and 22 patients, respectively). Triple-negative breast cancers are likely to be *BRCA1* mutation positive [83], while HER2 amplification is uncommon in *BRCA1* and *BRCA2* mutation carriers [84]. Although subgroup differences are noticeable, results still show that both subgroups had breakpoints on all chromosomes related to NPC (Figure S1A in Multimedia Appendix 5).

Tumor-Infiltrating Lymphocytes

Tumor-infiltrating lymphocytes (TILs) are biomarkers for predicting breast cancer prognosis [85,86]. To test whether TILs cause chromosome breaks, breakpoint numbers in 16 breast cancers with severe lymphocyte infiltration were compared to 17 breast cancers with nil lymphocyte infiltration. The 2-tailed Student *t* test could not reject the null hypothesis that the numbers of breakpoints were statistically identical (P=.70; Figure S1B in Multimedia Appendix 5). This result does not rule out differences in prognosis due to differences in lymphocyte infiltration.

Retroviruses

Retrovirus contributions to structural variations were estimated using data from cancer in 38 different tissues [87]. Retrotransposons make relatively modest contributions to breast cancer compared to, say, esophageal or oral (gums) cancer (Multimedia Appendix 5). EBV can transactivate endogenous retroviruses [11,87,88]. DNA near some breast cancer breakpoints resembles porcine endogenous retrovirus, HERV, and HIV1 (eg, Figure 5). The human genome also contains DNA matching the retrovirus mouse mammary tumor virus [89,90] at 23 sites that give BLAST homology scores>200. HPV variants are DNA viruses that are also implicated in breast cancer. HPVs were not assessed further, but they occasionally matched DNA near breast cancer breakpoints.

Common Fragile Sites

Common fragile sites are site-specific breaks seen on metaphase chromosomes after inhibiting DNA synthesis via DNA polymerase inhibitors. Some common fragile sites [54] aligned with breast cancer breaks on chromosome 1, but breakpoints on most other chromosomes were incompatible. Chromosomes 8, 9, 11 - 15, 17-19, 21, and 22 do not have common fragile sites but still have many breast cancer breaks [91]. However, the human genome has over 13 million palindromes that are ≤40 bp [92]. The generation of rare fragile sites by palindromes or their attraction to EBV cannot be excluded.

Imperfect Palindrome Repeats

An alternative explanation for EBV-related carcinogenesis involves the docking of EBNA1 virus-tethering protein at imperfect palindromes [93] tandemly repeated on chromosome 11. The docked EBNA1 binds EBV circular episomes, and chromosome 11 breaks initiate malignancy. To test this explanation, existing literature data were first compared to the specific human EBNA1-binding site [48,66,94]. The results (Table S2 in Multimedia Appendix 4) are incompatible with a single host sequence binding EBNA1.

BLAST analysis showed that matches to the imperfect palindrome were likely due to pure chance with E values between 16 and 964 for 4352 matches, from 12 to 18 bp. Chromosome 11 had only 197 of these 4352 matches, and none were near the palindromic region. The prototype DNA palindrome (Table S2 in Multimedia Appendix 4, line 2) produced 7074 matches with E values ranging from 0.25 to 964. Further BLAST analyses of the slightly different docking sequence in EBNA1-DNA crystals (Table S3 in Multimedia Appendix 4, line 1) against other genome assemblies [95]

revealed matches on chromosomes 2, 19, 4, and 12. Various isolates of HIVs had 52 matching sequences.

In 94 BL samples from patients who were EBV positive, breakpoints concentrated within chromosomes 2, 8, 13, 14, and 22 (Figure S1D in Multimedia Appendix 5). Chromosome 14 contained 610 breakpoints (*IgVH* regions), and chromosome 2 (*IgVK* regions) contained 522 breakpoints. EBV hijacks activation-induced cytidine deaminase, a mutagenic enzyme that generates antibody gene variants in response to myriad antigens. In the 94 EBV-positive BL cases, the palindromic locus was nearly 100 million bp away from the principal breakpoint coordinates (Figure S1E in Multimedia Appendix 5). Only 19 (20%) of the 94 patients who were EBV positive [96] had breakpoints anywhere on chromosome 11. The palindromic locus was also not involved in diverse cancers from 8227 patients [97] (Multimedia Appendix 5).

The results in this section show that alternative explanations that invoke subgroups, TILs, retroviruses, or a specific palindromic repeat locus are incompatible with the associations between EBV-associated and breast cancers .

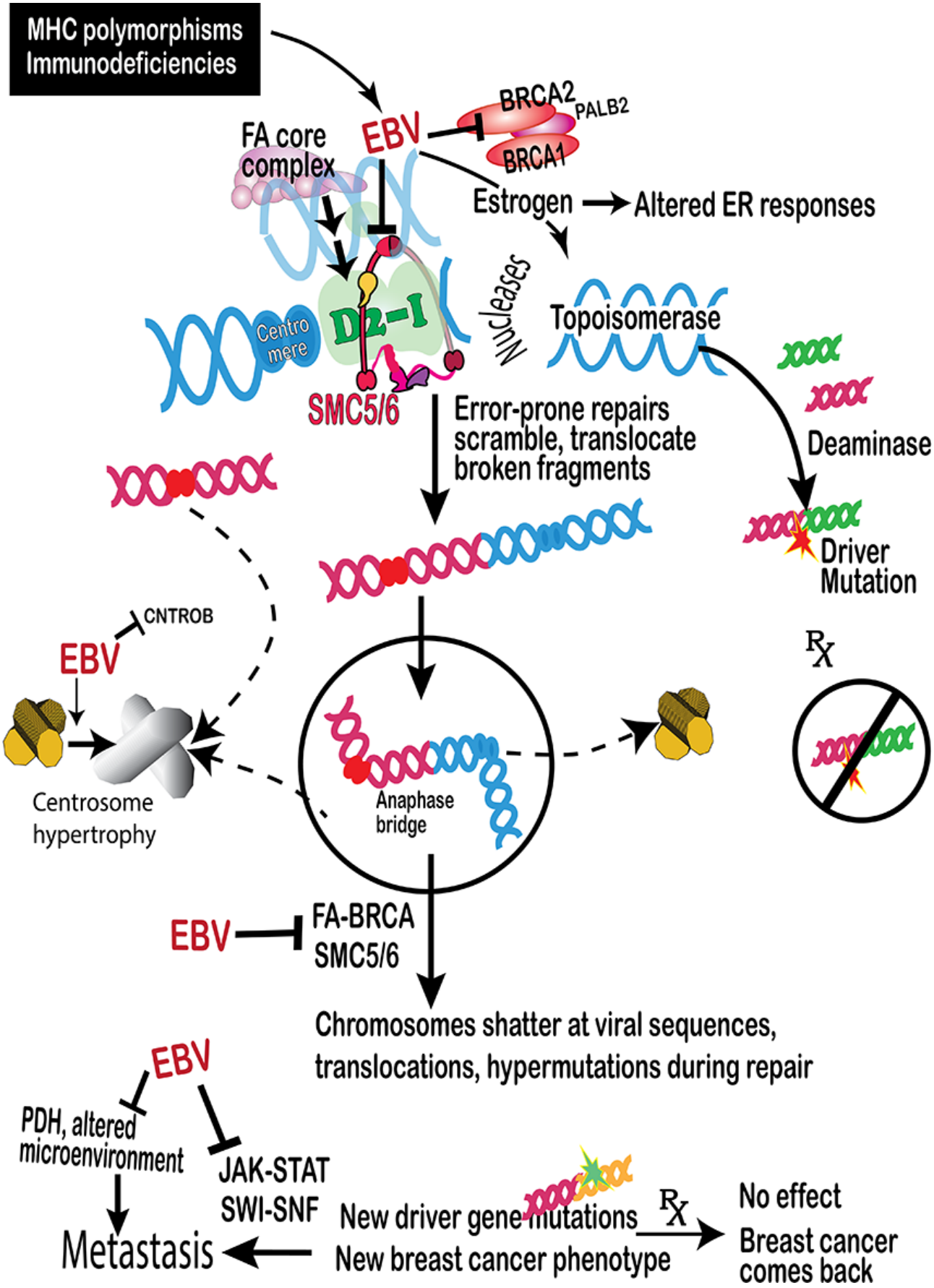
Discussion

Principal Findings

This study finds that EBV contributes to breast cancer by disabling safeguards against tumors. Cancer then occurs because the safeguards remain disabled even if the virus is cleared. Multiple independent analyses identified residual genetic and epigenetic damage in cancer genomes and formed the basis of the model in Figure 10. Breakpoints in breast cancers in high-risk women, sporadic breast cancers, and even ovarian cancers cluster around breakpoints in known EBV-related cancers, including NPC, BL, DLBCL, and GC. Some genes clustered near breakpoints in these diverse EBV-associated cancers are critical to preventing breast cancers. Some breast cancer breakpoints are near genes at EBV-docking sites. Varying numbers of DNA breaks occur within the highly polymorphic forms of MHC region genes on chromosome 6. This damage adds to susceptible polymorphisms and immunodeficiencies to help explain why not everyone develops EBV-related cancers. Near the MHC region on chromosome 6, piRNA sequences are regularly interspaced between viral DNA sequences. The sandwiched arrangements are presumptive evidence of past infection and probably represent a DNA defense mechanism. These defenses fail when chromosome 6 breaks apart near start points of the large number of repetitive viral sequences in the human genome. The viral sequences confuse repairs already damaged by EBV, and bursts of mutation occur where scrambled fragments ligate. EBV disables the most reliable restoration of broken chromosomes back to their native forms, so repairs form structures with multiple centromeres. These structures undergo additional rounds of fragmentation during cell division. The process continually forms new cancer driver mutations and allows cancer to come back after successful therapy (Figure 10). An EBV methylation signature on chromosome 6 was far more abundant in 1538 breast cancers than in normal controls. Finally, EBV facilitates its own replication by damaging JAK-STAT and SWI-SNF signaling pathways, which pushes breast cancer into metastasis, while virus-associated changes on chromosome 3p interfere with the destruction of metastatic cells. Models [8,98] of EBV-infected human mammary cell cultures transplanted into immunosuppressed mice and EBV loss from NPC cells are consistent with these results.

The study herein has current and future clinical implications in addressing cancers and chronic diseases. An early childhood vaccine against EBV may reduce the incidence of breast cancer on a global scale. If this vaccine even approaches the effectiveness of the HPV vaccine for cervical cancer, then the reduction of breast cancer incidence would be substantial. In breast cancer cases where active infection can be demonstrated, immunotherapy or antivirals can be considered. The results also heighten concern about hidden dangers from viral infections. EBV infection leaves behind persistent genome abnormalities (“long EBV”) linked to breast cancer. Not everyone develops an EBV-related cancer even though almost everyone is infected, suggesting risk assessment should include MHC polymorphisms. MHC genes have abundant connections to both EBV infection [99] and breast cancer [100-102].

Figure 10. Model proposed to explain the results. EBV causes serious disease in only some people due to MHC variants and other damage to the immune system. Viral nucleases are one source of chromosome breaks. EBV causes inappropriate expression of estrogen and transcription targets of occupied estrogen receptors. Transcription induced by artificially high estrogen levels then induces topoisomerase-mediated DNA breaks. EBV-mediated deregulation of estrogen production, topoisomerase activity, and deaminase activation then collaborate to cause chromosome breaks and drive translocations [68]. EBV-associated cancers share additional genome deficits with breast cancers, which interfere with restoring the genome from DNA crosslinks and DNA double-strand breaks. If crosslinks and DNA breaks persist during cell division, they also cause chromosome rearrangements and cancer. The cancer safeguards targeted by EBV extend to the *BRCA* pathway, FA proteins, an SMC5/6 scaffold, JAK-STAT signaling, and the SWI-SNF chromatin remodeling complex. EBV: Epstein-Barr virus; ER: estrogen receptor; FA: Fanconi anemia; MHC: major histocompatibility complex; PDH: pyruvate dehydrogenase; SWI-SNF: switch/sucrose non-fermentable.



The strategy of using bioinformatics to identify markers of "long EBV" may well work for other cancers, multiple sclerosis [103], and other chronic diseases that are currently unexplained. Testing for persistent viral damage in genomes from biopsies is a new method for screening for breast cancer risk. The results may inform further prevention and treatment decisions. Cancer

drug therapy has focused on finding and destroying cancer-driver gene products. The drugs are initially effective, sometimes for long periods, but then stop working. The cycles represented in Figure 10 are an occult, underlying process that can now be evaluated. Cancer treatment generates new clones that do not exist in the original population [104]. The underlying genome damage and EBV scars continually produce new cancer-driver mutations. Some antigens targeted by successful therapy for hematologic malignancies [105], such as DLBCL, may also be effective for breast cancers. The idea that breast cancers and hematologic malignancies can have similar breakpoints and translocation fusions suggests that there may be many more susceptible targets and that there are options to overcome resistance or tolerance [106]. The findings may further stimulate research into other EBV-associated diseases and cancers, leading to better and broader understanding.

Estrogen has been thought to generate the initial chromosome breakpoints leading to translocations in human breast cancer. However, young boys with BL do not produce estrogen from ovaries, yet Figure 3 shows that their malignant B-cells have many breakpoints [68,107] that approximately match breast cancer breakpoints. Normally, aromatase catalyzes the rate-limiting step in estrogen production [108], and aromatase acting on androgens is the primary source of most estrogens in breast tissue [109]. EBV-infected cells lose control of aromatase activity [108]. An EBV-mediated increase in aromatase activity explains why locations of breakpoints (Multimedia Appendix 5) are relatively independent of estrogen receptor status in breast cancer [68] and resemble locations in lymphoid cells (Figures 1-4 and 9). Transcription in response to artificially high estrogen levels created by EBV then induces topoisomerase-mediated DNA breaks. Double-strand break repair genes remove topoisomerase from these complexes, but damage to this process leaves pathological enzyme complexes still bound at a DNA breakpoint [110-112]. As shown in Figure 3, topoisomerase itself may be damaged. In either case, EBV-mediated deregulation of estrogen production, topoisomerase activity, and deaminases then collaborate to cause chromosome breaks and drive breast cancer.

Breast cancer chromosome breakpoints cluster around genes near EBV-binding sites (Figure 4), further suggesting that EBV participates in causing the breaks. The breaks lead to pathogenic chromosome rearrangements because EBV-induced damage forces restoration into error-prone methods by suppressing FA-BRCA pathway intermediates [14,15]. Repairs using the FA-BRCA pathway [113] need chromatin access, which requires the SMC5/6 cohesin complex [114,115]. In one scenario shown in Figure 10, SMC5/6 interacts with a crucial pathway intermediate, the FANCD2-FANCI heterodimer (“D2-I”) [17,116]. EBV variants deplete SMC5/6, preventing FA-BRCA–mediated DNA repairs and leading to chromosomes with too many centromeres. When mitosis pulls apart multicentromere chromosome structures, the forces shatter the chromosome and induce mutation storms [35]. EBV thus threatens a sprawling, interconnected repair system, including the BRCA pathway, FA proteins, an SMC5/6 scaffold, JAK-STAT signaling, and the SWI-SNF chromatin remodeling complex (Figure 10).

Of course, other environmental, genetic, or lifestyle factors also participate in breast cancer development, but EBV infection exacerbates their effects. Genome deficits in EBV-associated cancers and breast cancers interfere with restoring chromosomes from damage due to natural processes and exogenous mutagens. Some of this damage requires repair pathways that are subject to EBV interference.

Evidence underlying the model in Figure 10 has independent support from the literature. For example, viral load is a marker for the extent of cell-free DNA fragmentation [117]. EBV-mediated transformation routinely generates abnormal karyotypes [118]. The binding of EBNA1 sequence variants increases NPC risk and drives EBV lytic gene expression [119,120], which requires EBV-encoded nucleases [121-123]. Other herpesviruses related to EBV share the ability to fragment DNA and subvert DNA repair pathways [124-126]. EBV facilitates its own replication by interfering with signaling pathways that prevent metastasis [29,127-130]. Independent literature supports EBV participation in metastasis and the results shown in Figure 9. NPC has the highest metastatic rate among all head and neck cancers, and the levels of circulating EBV markers are highly predictive [10]. Finding EBV in lymph nodes of patients with NPC or primary cancer at an unknown site helps detect metastasis [131]. NPC patients with ≥500 copies of EBV per mL plasma had significantly higher rates of liver metastasis than patients with lower EBV levels [132]. EBV-infected B-cells and breast cancer cells both have amplified centrosomes (Figure 10), the mitosis-organizing centers that exert structural control over cell division. The EBV protein thymidine kinase takes up residence in the centrosome [133], and another EBV protein, BNRF1, initiates centrosome amplification in infected B-cells [134]. Overduplication of centrosomes confuses chromatid attachments to spindle fibers during mitosis. Chromosomes do not distribute properly into daughter cells, creating mistakes when the genome replicates [134,135]. Neither centrosome amplification nor chromosome fragmentation (chromothripsis) requires large numbers of viral particles or active infection.

Further bioinformatic tests may still add significant additional information. EBV activation brings massive changes to host chromatin methylation and structure [47,51,136]. Breast cancers have hundreds of these changes [34]. Results here further implicate epigenetic effects, so EBV effects on breast cancer epigenetics should be explored in more detail. EBV is implicated in cancers in multiple additional organs, and the methods developed here may help clarify its potential contributions. Predictions based on virus-human interaction structural biology may also be helpful. The ultimate direct test will be whether childhood recipients of an anti-EBV vaccine have reduced breast cancer incidence. If it even approaches the reduction of cervical cancer achieved by the HPV vaccine (up to 94%), a childhood EBV vaccine could effectively prevent many cases of breast cancer.

Limitations

EBV itself creates a limitation because the virus can disappear after causing pathogenic genome damage that allows breast cancer to develop. This transitory virus presence forces the use

of bioinformatics to look for persistent genome damage EBV leaves behind. EBV disappearance questions whether a group of cancers with EBV connections also contains “sporadic” cancers typed as EBV negative. The EBV-negative forms may have merely lost the criteria used to identify EBV infection, but EBV-related genome damage may still remain. Another limitation is that compared to breast cancers, known EBV-linked cancers such as GC, BL, and NPC are less common, so genome sequence data are also less common.

Conclusions

In summary, early childhood immunizations against inactivated EBV or selected EBV gene products may significantly reduce the incidence of breast, ovarian, and other cancers, and potentially unexplained chronic diseases. EBV variants lead to DNA breaks, mitotic abnormalities, and the loss of safeguards that protect against breast cancer and its metastasis. Breast cancer breakpoints cluster around breakpoints in EBV cancers, disrupting genes essential to prevent viral infection and breast cancers. A CRISPR-like region on chromosome 6 sequesters some of the thousands of pieces of EBV sequences in the human genome. The same area of chromosome 6 undergoes variable damage in breast cancer, contributing to the reason not everyone with EBV infection develops cancer. In susceptible people, EBV infection leaves behind pathogenic cancer-associated genome abnormalities (“long EBV”). Clinical implications include improvements in evaluating the chances that cancer will return, increased use of immunotherapy for patients with breast cancer that have active infection, and greater urgency in developing an effective EBV vaccine,.

Data Availability

The primary dataset and calculations that were generated or analyzed during this study are included. Datasets not included are freely available from the original sources or the corresponding author on reasonable request.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Glossary and abbreviations.
[DOCX File, 20 KB - xmed_v6i1e50712_app1.docx]

Multimedia Appendix 2
Most of the calculations used in this work.
[XLSX File, 7644 KB - xmed_v6i1e50712_app2.xlsx]

Multimedia Appendix 3
The exact distances between nasopharyngeal cancer and breast cancer breakpoints on chromosome 1. These breaks gather around a few low valleys that periodically occur across the whole chromosome, but the data points are too numerous to display, making the results difficult to interpret.
[PNG File, 213 KB - xmed_v6i1e50712_app3.png]

Multimedia Appendix 4
Gene functions at breast cancer breakpoints that clustered around breakpoints in EBV-associated cancers (GC, BL, and NPC), and EBNA1-binding sequences reported in the human genome. BL: Burkitt lymphoma; EBNA1: Epstein-Barr virus nuclear antigen 1; EBV: Epstein-Barr virus; NPC: nasopharyngeal cancer.
[DOCX File, 31 KB - xmed_v6i1e50712_app4.docx]

Multimedia Appendix 5
Alternative explanations.
[PNG File, 374 KB - xmed_v6i1e50712_app5.png]

References

1. DeSantis CE, Ma J, Gaudet MM, et al. Breast cancer statistics, 2019. CA Cancer J Clin 2019 Nov;69(6):438-451. [doi: 10.3322/caac.21583] [Medline: 31577379]
2. QuickStats: age-adjusted death rates* for female breast cancer,† by state — National Vital Statistics System, United States, 2019§. MMWR Morb Mortal Wkly Rep 2021 Oct 1;70(39):1391. [doi: 10.15585/mmwr.mm7039a6] [Medline: 34591833]
3. Sarid R, Gao SJ. Viruses and human cancer: from detection to causality. Cancer Lett 2011 Jun 28;305(2):218-227. [doi: 10.1016/j.canlet.2010.09.011] [Medline: 20971551]

4. zur Hausen H. Papillomaviruses and cancer: from basic studies to clinical application. Nat Rev Cancer 2002 May;2(5):342-350. [doi: 10.1038/nrc798] [Medline: 12044010]
5. Balfour HHJ, Sifakis F, Sliman JA, Knight JA, Schmeling DO, Thomas W. Age-specific prevalence of Epstein–Barr virus infection among individuals aged 6–19 years in the United States and factors affecting its acquisition. J Infect Dis 2013 Oct 15;208(8):1286-1293. [doi: 10.1093/infdis/jit321] [Medline: 23868878]
6. Huo Q, Zhang N, Yang Q. Epstein-Barr virus infection and sporadic breast cancer risk: a meta-analysis. PLoS One 2012;7(2):e31656. [doi: 10.1371/journal.pone.0031656] [Medline: 22363698]
7. Fina F, Romain S, Ouafik L, et al. Frequency and genome load of Epstein-Barr virus in 509 breast cancers from different geographical areas. Br J Cancer 2001 Mar 23;84(6):783-790. [doi: 10.1054/bjoc.2000.1672] [Medline: 11259092]
8. Hu H, Luo ML, Desmedt C, et al. Epstein-Barr virus infection of mammary epithelial cells promotes malignant transformation. EBioMedicine 2016 Jul;9:148-160. [doi: 10.1016/j.ebiom.2016.05.025] [Medline: 27333046]
9. God JM, Haque A. Burkitt lymphoma: pathogenesis and immune evasion. J Oncol 2010;2010:516047. [doi: 10.1155/2010/516047] [Medline: 20953370]
10. Xu M, Yao Y, Chen H, et al. Genome sequencing analysis identifies Epstein-Barr virus subtypes associated with high risk of nasopharyngeal carcinoma. Nat Genet 2019 Jul;51(7):1131-1136. [doi: 10.1038/s41588-019-0436-5] [Medline: 31209392]
11. Bruce JP, To KF, Lui VWY, et al. Whole-genome profiling of nasopharyngeal carcinoma reveals viral-host co-operation in inflammatory NF-κB activation and immune escape. Nat Commun 2021 Jul 7;12(1):4193. [doi: 10.1038/s41467-021-24348-6] [Medline: 34234122]
12. Fountzilas G, Psyrri A, Giannoulatou E, et al. Prevalent somatic BRCA1 mutations shape clinically relevant genomic patterns of nasopharyngeal carcinoma in Southeast Europe. Int J Cancer 2018 Jan 1;142(1):66-80. [doi: 10.1002/ijc.31023] [Medline: 28857155]
13. Devanaboyina M, Kaur J, Whiteley E, et al. NF-κB signaling in tumor pathways focusing on breast and ovarian cancer. Oncol Rev 2022 Oct 3;16:10568. [doi: 10.3389/or.2022.10568] [Medline: 36531159]
14. Lung RWM, Tong JHM, Ip LM, et al. EBV-encoded miRNAs can sensitize nasopharyngeal carcinoma to chemotherapeutic drugs by targeting BRCA1. J Cell Mol Med 2020 Nov;24(22):13523-13535. [doi: 10.1111/jcmm.16007] [Medline: 33074587]
15. Hau PM, Tsao SW. Epstein-Barr virus hijacks DNA damage response transducers to orchestrate its life cycle. Viruses 2017 Nov 16;9(11):341. [doi: 10.3390/v9110341] [Medline: 29144413]
16. Dheekollu J, Wiedmer A, Ayyanathan K, Deakyne JS, Messick TE, Lieberman PM. Cell-cycle-dependent EBNA1-DNA crosslinking promotes replication termination at oriP and viral episome maintenance. Cell 2021 Feb 4;184(3):643-654.e13. [doi: 10.1016/j.cell.2020.12.022] [Medline: 33482082]
17. Rossi F, Helbling-Leclerc A, Kawasumi R, et al. SMC5/6 acts jointly with Fanconi anemia factors to support DNA repair and genome stability. EMBO Rep 2020 Feb 5;21(2):e48222. [doi: 10.15252/embr.201948222] [Medline: 31867888]
18. Yiu SPT, Guo R, Zerbe C, Weekes MP, Gewurz BE. Epstein-Barr virus BNRF1 destabilizes SMC5/6 cohesin complexes to evade its restriction of replication compartments. Cell Rep 2022 Mar 8;38(10):110411. [doi: 10.1016/j.celrep.2022.110411] [Medline: 35263599]
19. Fan Y, Ying H, Wu X, et al. The mutational pattern of homologous recombination (HR)-associated genes and its relevance to the immunotherapeutic response in gastric cancer. Cancer Biol Med 2020 Nov 15;17(4):1002-1013. [doi: 10.20892/j.issn.2095-3941.2020.0089] [Medline: 33299649]
20. Parvin S, Labrada AR, Santiago GE, et al. Novel role of LMO2 in DNA repair control in diffuse large B cell lymphoma. Blood 2016 Dec 2;128(22):776-776. [doi: 10.1182/blood.V128.22.776.776]
21. Busch K, Keller T, Fuchs U, et al. Identification of two distinct MYC breakpoint clusters and their association with various IGH breakpoint regions in the t(8;14) translocations in sporadic Burkitt-lymphoma. Leukemia 2007 Aug;21(8):1739-1751. [doi: 10.1038/sj.leu.2404753] [Medline: 17541401]
22. Chong LC, Ben-Neriah S, Slack GW, et al. High-resolution architecture and partner genes of MYC rearrangements in lymphoma with DLBCL morphology. Blood Adv 2018 Oct 23;2(20):2755-2765. [doi: 10.1182/bloodadvances.2018023572] [Medline: 30348671]
23. Xu J, Chen Y, Olopade OI. MYC and breast cancer. Genes Cancer 2010 Jun;1(6):629-640. [doi: 10.1177/1947601910378691] [Medline: 21779462]
24. Umbreit NT, Zhang CZ, Lynch LD, et al. Mechanisms generating cancer genome complexity from a single cell division error. Science 2020 Apr 17;368(6488):eaba0712. [doi: 10.1126/science.aba0712] [Medline: 32299917]
25. McClintock B. The stability of broken ends of chromosomes in Zea mays. Genetics 1941 Mar;26(2):234-282. [doi: 10.1093/genetics/26.2.234] [Medline: 17247004]
26. Lee JJK, Jung YL, Cheong TC, et al. ERα-associated translocations underlie oncogene amplifications in breast cancer. Nature New Biol 2023 Jun;618(7967):1024-1032. [doi: 10.1038/s41586-023-06057-w] [Medline: 37198482]
27. McAulay KA, Higgins CD, Macsween KF, et al. HLA class I polymorphisms are associated with development of infectious mononucleosis upon primary EBV infection. J Clin Invest 2007 Oct;117(10):3042-3048. [doi: 10.1172/JCI32377] [Medline: 17909631]
28. Aboulaghras S, Khalid A, Makeen HA, et al. Polymorphism of HLA and susceptibility of breast cancer. Front Biosci (Landmark Ed) 2024 Feb 5;29(2):55. [doi: 10.31083/j.fbl2902055] [Medline: 38420797]

29. Yates LR, Knappskog S, Wedge D, et al. Genomic evolution of breast cancer metastasis and relapse. Cancer Cell 2017 Aug 14;32(2):169-184.e7. [doi: 10.1016/j.ccell.2017.07.005] [Medline: 28810143]
30. Friedenson B. Dewey defeats Truman and cancer statistics. J Natl Cancer Inst 2009 Aug 19;101(16):1157. [doi: 10.1093/jnci/djp203] [Medline: 19561316]
31. Nik-Zainal S, Davies H, Staaf J, et al. Landscape of somatic mutations in 560 breast cancer whole-genome sequences. Nature New Biol 2016 Jun 2;534(7605):47-54. [doi: 10.1038/nature17676] [Medline: 27135926]
32. Staaf J, Glodzik D, Bosch A, et al. Whole-genome sequencing of triple-negative breast cancers in a population-based clinical study. Nat Med 2019 Oct;25(10):1526-1533. [doi: 10.1038/s41591-019-0582-4] [Medline: 31570822]
33. Nones K, Johnson J, Newell F, et al. Whole-genome sequencing reveals clinically relevant insights into the aetiology of familial breast cancers. Ann Oncol 2019 Jul 1;30(7):1071-1079. [doi: 10.1093/annonc/mdz132] [Medline: 31090900]
34. Batra RN, Lifshitz A, Vidakovic AT, et al. DNA methylation landscapes of 1538 breast cancers reveal a replication-linked clock, epigenomic instability and cis-regulation. Nat Commun 2021 Sep 13;12(1):5406. [doi: 10.1038/s41467-021-25661-w] [Medline: 34518533]
35. Cortés-Ciriano I, Lee JJK, Xi R, et al. Comprehensive analysis of chromothripsis in 2,658 human cancers using whole-genome sequencing. Nat Genet 2020 Mar;52(3):331-341. [doi: 10.1038/s41588-019-0576-7] [Medline: 32025003]
36. Petrucelli N, Daly MB, Pal T. BRCA1- and BRCA2-associated hereditary breast and ovarian cancer. In: Adam MP, Mirzaa GM, Pagon RA, Wallace SE, Bean LJH, Gripp KW, editors. GeneReviews: University of Washington, Seattle; 1993. URL: https://www.ncbi.nlm.nih.gov/books/NBK1247/ [accessed 2025-01-09]
37. Lalloo F, Varley J, Moran A, et al. BRCA1, BRCA2 and TP53 mutations in very early-onset breast cancer with associated risks to relatives. Eur J Cancer 2006 May;42(8):1143-1150. [doi: 10.1016/j.ejca.2005.11.032] [Medline: 16644204]
38. Joos S, Falk MH, Lichter P, et al. Variable breakpoints in Burkitt lymphoma cells with chromosomal t(8;14) translocation separate c-myc and the IgH locus up to several hundred kb. Hum Mol Genet 1992 Nov;1(8):625-632. [doi: 10.1093/hmg/1.8.625] [Medline: 1301171]
39. Joos S, Haluska FG, Falk MH, et al. Mapping chromosomal breakpoints of Burkitt's t(8;14) translocations far upstream of c-myc. Cancer Res 1992 Dec 1;52(23):6547-6552. [Medline: 1330296]
40. López C, Kleinheinz K, Aukema SM, et al. Genomic and transcriptomic changes complement each other in the pathogenesis of sporadic Burkitt lymphoma. Nat Commun 2019 Mar 29;10(1):1459. [doi: 10.1038/s41467-019-08578-3] [Medline: 30926794]
41. Xing R, Zhou Y, Yu J, et al. Whole-genome sequencing reveals novel tandem-duplication hotspots and a prognostic mutational signature in gastric cancer. Nat Commun 2019 May 2;10(1):2037. [doi: 10.1038/s41467-019-09644-6] [Medline: 31048690]
42. Nangalia J, Campbell PJ. Genome sequencing during a patient's journey through cancer. N Engl J Med 2019 Nov 28;381(22):2145-2156. [doi: 10.1056/NEJMra1910138] [Medline: 31774959]
43. Mount DW. Using the Basic Local Alignment Search Tool (BLAST). CSH Protoc 2007 Jul 1;2007(7):pdb.top17. [doi: 10.1101/pdb.top17] [Medline: 21357135]
44. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol 1990 Oct 5;215(3):403-410. [doi: 10.1016/S0022-2836(05)80360-2] [Medline: 2231712]
45. Zhang Z, Schwartz S, Wagner L, Miller W. A greedy algorithm for aligning DNA sequences. J Comput Biol 2000;7(1-2):203-214. [doi: 10.1089/10665270050081478] [Medline: 10890397]
46. Rangwala SH, Kuznetsov A, Ananiev V, et al. Accessing NCBI data using the NCBI Sequence Viewer and Genome Data Viewer (GDV). Genome Res 2021 Jan;31(1):159-169. [doi: 10.1101/gr.266932.120] [Medline: 33239395]
47. Kim KD, Tanizawa H, de Leo A, et al. Epigenetic specifications of host chromosome docking sites for latent Epstein-Barr virus. Nat Commun 2020 Feb 13;11(1):877. [doi: 10.1038/s41467-019-14152-8] [Medline: 32054837]
48. Lu F, Wikramasinghe P, Norseen J, et al. Genome-wide analysis of host-chromosome binding sites for Epstein-Barr virus nuclear antigen 1 (EBNA1). Virol J 2010 Oct 7;7:262. [doi: 10.1186/1743-422X-7-262] [Medline: 20929547]
49. Sai Lakshmi S, Agrawal S. piRNABank: a web resource on classified and clustered Piwi-interacting RNAs. Nucleic Acids Res 2008 Jan;36(Database issue):D173-D177. [doi: 10.1093/nar/gkm696] [Medline: 17881367]
50. Wang J, Zhang P, Lu Y, et al. piRBase: a comprehensive database of piRNA sequences. Nucleic Acids Res 2019 Jan 8;47(D1):D175-D180. [doi: 10.1093/nar/gky1043] [Medline: 30371818]
51. Tang MH, Varadan V, Kamalakaran S, Zhang MQ, Dimitrova N, Hicks J. Major chromosomal breakpoint intervals in breast cancer co-localize with differentially methylated regions. Front Oncol 2012 Dec 27;2:197. [doi: 10.3389/fonc.2012.00197] [Medline: 23293768]
52. Mann HB, Whitney DR. On a test of whether one of two random variables is stochastically larger than the other. Ann Math Statist 1947 Mar;18(1):50-60. [doi: 10.1214/aoms/1177730491]
53. Shapiro SS, Wilk MB. An analysis of variance test for normality (complete samples). Biometrika 1965 Dec;52(3/4):591-611. [doi: 10.2307/2333709]
54. Kumar R, Nagpal G, Kumar V, Usmani SS, Agrawal P, Raghava GPS. HumCFS: a database of fragile sites in human chromosomes. BMC Genomics 2019 Apr 18;19(Suppl 9):985. [doi: 10.1186/s12864-018-5330-5] [Medline: 30999860]

55. Maccaroni K, Balzano E, Mirimao F, Giunta S, Pelliccia F. Impaired replication timing promotes tissue-specific expression of common fragile sites. Genes (Basel) 2020 Mar 19;11(3):326. [doi: 10.3390/genes11030326] [Medline: 32204553]
56. European Commission: Directorate-General for Research and Innovation. Improving access to and reuse of research results, publications and data for scientific purposes – study to evaluate the effects of the EU copyright framework on research and the effects of potential interventions and to identify and present relevant provisions for research in EU data and digital legislation, with a focus on rights and obligations. : Publications Office of the European Union; 2024 URL: https://data.europa.eu/doi/10.2777/633395 [accessed 2025-01-09]
57. Zheng B, Liu XL, Fan R, et al. The landscape of cell-free HBV integrations and mutations in cirrhosis and hepatocellular carcinoma patients. Clin Cancer Res 2021 Jul 1;27(13):3772-3783. [doi: 10.1158/1078-0432.CCR-21-0002] [Medline: 33947693]
58. Foster KA, Harrington P, Kerr J, et al. Somatic and germline mutations of the BRCA2 gene in sporadic ovarian cancer. Cancer Res 1996 Aug 15;56(16):3622-3625. [Medline: 8705994]
59. Friedenson B. The BRCA1/2 pathway prevents hematologic cancers in addition to breast and ovarian cancers. BMC Cancer 2007 Aug 6;7:152. [doi: 10.1186/1471-2407-7-152] [Medline: 17683622]
60. Friedenson B. Comment on 'The incidence of leukaemia in women with BRCA1 and BRCA2 mutations: an International Prospective Cohort Study'. Br J Cancer 2016 Aug 23;115(5):e2. [doi: 10.1038/bjc.2016.192] [Medline: 27459694]
61. Parvin S, Ramirez-Labrada A, Aumann S, et al. LMO2 confers synthetic lethality to PARP inhibition in DLBCL. Cancer Cell 2019 Sep 16;36(3):237-249.e6. [doi: 10.1016/j.ccell.2019.07.007] [Medline: 31447348]
62. Compagno M, Lim WK, Grunn A, et al. Mutations of multiple genes cause deregulation of NF-kappaB in diffuse large B-cell lymphoma. Nature 2009 Jun 4;459(7247):717-721. [doi: 10.1038/nature07968] [Medline: 19412164]
63. Ceribelli M, Kelly PN, Shaffer AL, et al. Blockade of oncogenic IκB kinase activity in diffuse large B-cell lymphoma by bromodomain and extraterminal domain protein inhibitors. Proc Natl Acad Sci U S A 2014 Aug 5;111(31):11365-11370. [doi: 10.1073/pnas.1411701111] [Medline: 25049379]
64. Montes-Moreno S, Odqvist L, Diaz-Perez JA, et al. EBV-positive diffuse large B-cell lymphoma of the elderly is an aggressive post-germinal center B-cell neoplasm characterized by prominent nuclear factor-kB activation. Mod Pathol 2012 Jul;25(7):968-982. [doi: 10.1038/modpathol.2012.52] [Medline: 22538516]
65. Gröbner SN, Worst BC, Weischenfeldt J, et al. The landscape of genomic alterations across childhood cancers. Nature 2018 Mar 15;555(7696):321-327. [doi: 10.1038/nature25480] [Medline: 29489754]
66. Li JSZ, Abbasi A, Kim DH, Lippman SM, Alexandrov LB, Cleveland DW. Chromosomal fragile site breakage by EBV-encoded EBNA1 at clustered repeats. Nature 2023 Apr;616(7957):504-509. [doi: 10.1038/s41586-023-05923-x] [Medline: 37046091]
67. Hui KF, Chan TF, Yang W, et al. High risk Epstein‐Barr virus variants characterized by distinct polymorphisms in the EBER locus are strongly associated with nasopharyngeal carcinoma. Int J Cancer 2019 Jun 15;144(12):3031-3042. [doi: 10.1002/ijc.32049] [Medline: 30536939]
68. Friedenson B. Evidence of lesions from Epstein-Barr virus infection in human breast cancer genomes. medRxiv. Preprint posted online on Jun 26, 2024. [doi: 10.1101/2024.06.24.24309410]
69. Nurk S, Koren S, Rhie A, et al. The complete sequence of a human genome. Science 2022 Apr;376(6588):44-53. [doi: 10.1126/science.abj6987] [Medline: 35357919]
70. Luo WJ, Takakuwa T, Ham MF, et al. Epstein-Barr virus is integrated between REL and BCL-11A in American Burkitt lymphoma cell line (NAB-2). Lab Invest 2004 Sep;84(9):1193-1199. [doi: 10.1038/labinvest.3700152] [Medline: 15241441]
71. Tsao SW, Tsang CM, Lo KW. Epstein-Barr virus infection and nasopharyngeal carcinoma. Philos Trans R Soc Lond B Biol Sci 2017 Oct 19;372(1732):20160270. [doi: 10.1098/rstb.2016.0270] [Medline: 28893937]
72. Scott RS. Epstein-Barr virus: a master epigenetic manipulator. Curr Opin Virol 2017 Oct;26:74-80. [doi: 10.1016/j.coviro.2017.07.017] [Medline: 28780440]
73. Schott G, Garcia-Blanco MA. MHC class III RNA binding proteins and immunity. RNA Biol 2021 May;18(5):640-646. [doi: 10.1080/15476286.2020.1860388] [Medline: 33280511]
74. Zhou D, Lai M, Luo A, Yu CY. An RNA metabolism and surveillance quartet in the major histocompatibility complex. Cells 2019 Aug 30;8(9):1008. [doi: 10.3390/cells8091008] [Medline: 31480283]
75. Fernandes MT, Dejardin E, dos Santos NR. Context-dependent roles for lymphotoxin-β receptor signaling in cancer development. Biochim Biophys Acta 2016 Apr;1865(2):204-219. [doi: 10.1016/j.bbcan.2016.02.005] [Medline: 26923876]
76. Shiina T, Blancher A, Inoko H, Kulski JK. Comparative genomics of the human, macaque and mouse major histocompatibility complex. Immunology 2017 Feb;150(2):127-138. [doi: 10.1111/imm.12624] [Medline: 27395034]
77. Aissani B, Boehme AK, Wiener HW, Shrestha S, Jacobson LP, Kaslow RA. SNP screening of central MHC-identified HLA-DMB as a candidate susceptibility gene for HIV-related Kaposi's sarcoma. Genes Immun 2014 Sep;15(6):424-429. [doi: 10.1038/gene.2014.42] [Medline: 25008864]
78. Zhuang J, Huo Q, Yang F, Xie N. Perspectives on the role of histone modification in breast cancer progression and the advanced technological tools to study epigenetic determinants of metastasis. Front Genet 2020 Oct 29;11:603552. [doi: 10.3389/fgene.2020.603552] [Medline: 33193750]

79. Bayona-Feliu A, Barroso S, Muñoz S, Aguilera A. The SWI/SNF chromatin remodeling complex helps resolve R-loop-mediated transcription–replication conflicts. Nat Genet 2021 Jul;53(7):1050-1063. [doi: 10.1038/s41588-021-00867-2] [Medline: 33986538]
80. Harrod A, Lane KA, Downs JA. The role of the SWI/SNF chromatin remodelling complex in the response to DNA double strand breaks. DNA Repair (Amst) 2020 Sep;93:102919. [doi: 10.1016/j.dnarep.2020.102919] [Medline: 33087260]
81. Andrades A, Peinado P, Alvarez-Perez JC, et al. SWI/SNF complexes in hematological malignancies: biological implications and therapeutic opportunities. Mol Cancer 2023 Feb 21;22(1):39. [doi: 10.1186/s12943-023-01736-8] [Medline: 36810086]
82. Lu J. The Warburg metabolism fuels tumor metastasis. Cancer Metastasis Rev 2019 Jun;38(1-2):157-164. [doi: 10.1007/s10555-019-09794-5] [Medline: 30997670]
83. Chen H, Wu J, Zhang Z, et al. Association between BRCA status and triple-negative breast cancer: a meta-analysis. Front Pharmacol 2018 Aug 21;9:909. [doi: 10.3389/fphar.2018.00909] [Medline: 30186165]
84. Evans DG, Lalloo F, Howell S, Verhoef S, Woodward ER, Howell A. Low prevalence of HER2 positivity amongst BRCA1 and BRCA2 mutation carriers and in primary BRCA screens. Breast Cancer Res Treat 2016 Feb;155(3):597-601. [doi: 10.1007/s10549-016-3697-z] [Medline: 26888723]
85. Locy H, Verhulst S, Cools W, et al. Assessing tumor-infiltrating lymphocytes in breast cancer: a proposal for combining immunohistochemistry and gene expression analysis to refine scoring. Front Immunol 2022 Feb 11;13:794175. [doi: 10.3389/fimmu.2022.794175] [Medline: 35222378]
86. Takada K, Kashiwagi S, Asano Y, et al. Prediction of distant metastatic recurrence by tumor-infiltrating lymphocytes in hormone receptor-positive breast cancer. BMC Womens Health 2021 May 29;21(1):225. [doi: 10.1186/s12905-021-01373-7] [Medline: 34051785]
87. Meier UC, Cipian RC, Karimi A, Ramasamy R, Middeldorp JM. Cumulative roles for Epstein-Barr virus, human endogenous retroviruses, and human herpes virus-6 in driving an inflammatory cascade underlying MS pathogenesis. Front Immunol 2021 Nov 1;12:757302. [doi: 10.3389/fimmu.2021.757302] [Medline: 34790199]
88. Mameli G, Poddighe L, Mei A, et al. Expression and activation by Epstein Barr virus of human endogenous retroviruses-W in blood cells and astrocytes: inference for multiple sclerosis. PLoS One 2012;7(9):e44991. [doi: 10.1371/journal.pone.0044991] [Medline: 23028727]
89. Lawson JS, Glenn WK. Evidence for a causal role by mouse mammary tumour-like virus in human breast cancer. NPJ Breast Cancer 2019 Nov 7;5:40. [doi: 10.1038/s41523-019-0136-4] [Medline: 31728407]
90. Dudley JP, Golovkina TV, Ross SR. Lessons learned from mouse mammary tumor virus in animal models. ILAR J 2016;57(1):12-23. [doi: 10.1093/ilar/ilv044] [Medline: 27034391]
91. Li Y, Roberts ND, Wala JA, et al. Patterns of somatic structural variation in human cancer genomes. Nature 2020 Feb;578(7793):112-121. [doi: 10.1038/s41586-019-1913-9] [Medline: 32025012]
92. Ganapathiraju MK, Subramanian S, Chaparala S, Karunakaran KB. A reference catalog of DNA palindromes in the human genome and their variations in 1000 Genomes. Hum Genome Var 2020 Nov 20;7(1):40. [doi: 10.1038/s41439-020-00127-5] [Medline: 33298903]
93. Rawlins DR, Milman G, Hayward SD, Hayward GS. Sequence-specific DNA binding of the Epstein-Barr virus nuclear antigen (EBNA-1) to clustered sites in the plasmid maintenance region. Cell 1985 Oct;42(3):859-868. [doi: 10.1016/0092-8674(85)90282-x] [Medline: 2996781]
94. Bochkarev A, Barwell JA, Pfuetzner RA, Bochkareva E, Frappier L, Edwards AM. Crystal structure of the DNA-binding domain of the Epstein-Barr virus origin-binding protein, EBNA1, bound to DNA. Cell 1996 Mar 8;84(5):791-800. [doi: 10.1016/s0092-8674(00)81056-9] [Medline: 8625416]
95. Altemose N, Logsdon GA, Bzikadze AV, et al. Complete genomic and epigenetic maps of human centromeres. Science 2022 Apr;376(6588):eabl4178. [doi: 10.1126/science.abl4178] [Medline: 35357911]
96. Grande BM, Gerhard DS, Jiang A, et al. Genome-wide discovery of somatic coding and noncoding mutations in pediatric endemic and sporadic Burkitt lymphoma. Blood 2019 Mar 21;133(12):1313-1324. [doi: 10.1182/blood-2018-09-871418] [Medline: 30617194]
97. Kim TM, Xi R, Luquette LJ, Park RW, Johnson MD, Park PJ. Functional genomic analysis of chromosomal aberrations in a compendium of 8000 cancer genomes. Genome Res 2013 Feb;23(2):217-227. [doi: 10.1101/gr.140301.112] [Medline: 23132910]
98. Dittmer DP, Hilscher CJ, Gulley ML, Yang EV, Chen M, Glaser R. Multiple pathways for Epstein-Barr virus episome loss from nasopharyngeal carcinoma. Int J Cancer 2008 Nov 1;123(9):2105-2112. [doi: 10.1002/ijc.23685] [Medline: 18688856]
99. Li Q, Cohen JI. Epstein-Barr virus and the human leukocyte antigen complex. Curr Clin Microbiol Rep 2019 Sep;6(3):175-181. [doi: 10.1007/s40588-019-00120-9] [Medline: 33094090]
100. Lei PJ, Pereira ER, Andersson P, et al. Cancer cell plasticity and MHC-II–mediated immune tolerance promote breast cancer metastasis to lymph nodes. J Exp Med 2023 Sep 4;220(9):e20221847. [doi: 10.1084/jem.20221847] [Medline: 37341991]
101. Park IA, Hwang SH, Song IH, et al. Expression of the MHC class II in triple-negative breast cancer is associated with tumor-infiltrating lymphocytes and interferon signaling. PLoS One 2017 Aug 17;12(8):e0182786. [doi: 10.1371/journal.pone.0182786] [Medline: 28817603]

102. Song IH, Kim YA, Heo SH, et al. The association of estrogen receptor activity, interferon signaling, and MHC class I expression in breast cancer. Cancer Res Treat 2022 Oct;54(4):1111-1120. [doi: 10.4143/crt.2021.1017] [Medline: 34942685]
103. Bjornevik K, Cortese M, Healy BC, et al. Longitudinal analysis reveals high prevalence of Epstein-Barr virus associated with multiple sclerosis. Science 2022 Jan 21;375(6578):296-301. [doi: 10.1126/science.abj8222] [Medline: 35025605]
104. Aissa AF, Islam ABMMK, Ariss MM, et al. Single-cell transcriptional changes associated with drug tolerance and response to combination therapies in cancer. Nat Commun 2021 Mar 12;12(1):1628. [doi: 10.1038/s41467-021-21884-z] [Medline: 33712615]
105. Cao Y, Efetov SK, He M, et al. Updated clinical perspectives and challenges of chimeric antigen receptor-T cell therapy in colorectal cancer and invasive breast cancer. Arch Immunol Ther Exp (Warsz) 2023 Aug 11;71(1):19. [doi: 10.1007/s00005-023-00684-x] [Medline: 37566162]
106. Upton R, Banuelos A, Feng D, et al. Combining CD47 blockade with trastuzumab eliminates HER2-positive breast cancer cells and overcomes trastuzumab tolerance. Proc Natl Acad Sci U S A 2021 Jul 20;118(29):e2026849118. [doi: 10.1073/pnas.2026849118] [Medline: 34257155]
107. Burkhardt B, Michgehl U, Rohde J, et al. Clinical relevance of molecular characteristics in Burkitt lymphoma differs according to age. Nat Commun 2022 Jul 6;13(1):3881. [doi: 10.1038/s41467-022-31355-8] [Medline: 35794096]
108. Dochi H, Kondo S, Murata T, et al. Estrogen induces the expression of EBV lytic protein ZEBRA, a marker of poor prognosis in nasopharyngeal carcinoma. Cancer Sci 2022 Aug;113(8):2862-2877. [doi: 10.1111/cas.15440] [Medline: 35633182]
109. Zhu BT, Conney AH. Functional role of estrogen metabolism in target cells: review and perspectives. Carcinogenesis 1998 Jan 1;19(1):1-27. [doi: 10.1093/carcin/19.1.1] [Medline: 9472688]
110. Pommier Y, Nussenzweig A, Takeda S, Austin C. Human topoisomerases and their roles in genome stability and organization. Nat Rev Mol Cell Biol 2022 Jun;23(6):407-427. [doi: 10.1038/s41580-022-00452-3] [Medline: 35228717]
111. Sasanuma H, Tsuda M, Morimoto S, et al. BRCA1 ensures genome integrity by eliminating estrogen-induced pathological topoisomerase II-DNA complexes. Proc Natl Acad Sci U S A 2018 Nov 6;115(45):E10642-E10651. [doi: 10.1073/pnas.1803177115] [Medline: 30352856]
112. Manguso N, Kim M, Joshi N, et al. TDP2 is a regulator of estrogen-responsive oncogene expression. NAR Cancer 2024 Apr 8;6(2):zcae016. [doi: 10.1093/narcan/zcae016] [Medline: 38596431]
113. de Piccoli G, Cortes-Ledesma F, Ira G, et al. Smc5-Smc6 mediate DNA double-strand-break repair by promoting sister-chromatid recombination. Nat Cell Biol 2006 Sep;8(9):1032-1034. [doi: 10.1038/ncb1466] [Medline: 16892052]
114. Tapia-Alveal C, Lin SJ, O'Connell MJ. Functional interplay between cohesin and Smc5/6 complexes. Chromosoma 2014 Oct;123(5):437-445. [doi: 10.1007/s00412-014-0474-9] [Medline: 24981336]
115. Irwan ID, Cullen BR. The SMC5/6 complex: an emerging antiviral restriction factor that can silence episomal DNA. PLoS Pathog 2023 Mar 2;19(3):e1011180. [doi: 10.1371/journal.ppat.1011180] [Medline: 36862666]
116. Agashe S, Joseph CR, Reyes TAC, et al. Smc5/6 functions with Sgs1-Top3-Rmi1 to complete chromosome replication at natural pause sites. Nat Commun 2021 Apr 8;12(1):2111. [doi: 10.1038/s41467-021-22217-w] [Medline: 33833229]
117. Truszewska A, Wirkowska A, Gala K, et al. EBV load is associated with cfDNA fragmentation and renal damage in SLE patients. Lupus 2021 Jul;30(8):1214-1225. [doi: 10.1177/09612033211010339] [Medline: 33866897]
118. Volleth M, Zenker M, Joksic I, Liehr T. Long-term culture of EBV-induced human lymphoblastoid cell lines reveals chromosomal instability. J Histochem Cytochem 2020 Apr;68(4):239-251. [doi: 10.1369/0022155420910113] [Medline: 32108534]
119. Dheekollu J, Malecka K, Wiedmer A, et al. Carcinoma-risk variant of EBNA1 deregulates Epstein-Barr virus episomal latency. Oncotarget 2017 Jan 31;8(5):7248-7264. [doi: 10.18632/oncotarget.14540] [Medline: 28077791]
120. Buisson M, Géoui T, Flot D, et al. A bridge crosses the active-site canyon of the Epstein-Barr virus nuclease with DNase and RNase activities. J Mol Biol 2009 Aug 28;391(4):717-728. [doi: 10.1016/j.jmb.2009.06.034] [Medline: 19538972]
121. Wu CC, Liu MT, Chang YT, et al. Epstein-Barr virus DNase (BGLF5) induces genomic instability in human epithelial cells. Nucleic Acids Res 2010 Apr;38(6):1932-1949. [doi: 10.1093/nar/gkp1169] [Medline: 20034954]
122. Chiu SH, Wu MC, Wu CC, et al. Epstein-Barr virus BALF3 has nuclease activity and mediates mature virion production during the lytic cycle. J Virol 2014 May;88(9):4962-4975. [doi: 10.1128/JVI.00063-14] [Medline: 24554665]
123. Chiu SH, Wu CC, Fang CY, et al. Epstein-Barr virus BALF3 mediates genomic instability and progressive malignancy in nasopharyngeal carcinoma. Oncotarget 2014 Sep 30;5(18):8583-8601. [doi: 10.18632/oncotarget.2323] [Medline: 25261366]
124. O'Neill FJ, Miles CP. Chromosome changes in human cells induced by herpes simplex, types 1 and 2. Nature 1969 Aug 23;223(5208):851-852. [doi: 10.1038/223851a0] [Medline: 4307971]
125. Mertens ME, Knipe DM. Herpes simplex virus 1 manipulates host cell antiviral and proviral DNA damage responses. mBio 2021 Feb 9;12(1):e03552-20. [doi: 10.1128/mBio.03552-20] [Medline: 33563816]
126. Schumacher AJ, Mohni KN, Kan Y, Hendrickson EA, Stark JM, Weller SK. The HSV-1 exonuclease, UL12, stimulates recombination by a single strand annealing mechanism. PLoS Pathog 2012;8(8):e1002862. [doi: 10.1371/journal.ppat.1002862] [Medline: 22912580]
127. Ezeonwumelu IJ, Garcia-Vidal E, Ballana E. JAK-STAT pathway: a novel target to tackle viral infections. Viruses 2021 Nov 27;13(12):2379. [doi: 10.3390/v13122379] [Medline: 34960648]

128. Jangra S, Bharti A, Lui WY, et al. Suppression of JAK-STAT signaling by Epstein-Barr virus tegument protein BGLF2 through recruitment of SHP1 phosphatase and promotion of STAT2 degradation. J Virol 2021 Sep 27;95(20):e0102721. [doi: 10.1128/JVI.01027-21] [Medline: 34319780]
129. Wu DY, Krumm A, Schubach WH. Promoter-specific targeting of human SWI-SNF complex by Epstein-Barr virus nuclear protein 2. J Virol 2000 Oct;74(19):8893-8903. [doi: 10.1128/jvi.74.19.8893-8903.2000] [Medline: 10982332]
130. Su MT, Wang YT, Chen YJ, Lin SF, Tsai CH, Chen MR. The SWI/SNF chromatin regulator BRG1 modulates the transcriptional regulatory activity of the Epstein-Barr virus DNA polymerase processivity factor BMRF1. J Virol 2017 Apr 13;91(9):e02114-16. [doi: 10.1128/JVI.02114-16] [Medline: 28228591]
131. Nakao K, Yuge T, Mochiki M, Nibu KI, Sugasawa M. Detection of Epstein-Barr virus in metastatic lymph nodes of patients with nasopharyngeal carcinoma and a primary unknown carcinoma. Arch Otolaryngol Head Neck Surg 2003 Mar;129(3):338-340. [doi: 10.1001/archotol.129.3.338] [Medline: 12622545]
132. Tao D, Zhang N, Huang Q, et al. Association of Epstein-Barr virus infection with peripheral immune parameters and clinical outcome in advanced nasopharyngeal carcinoma. Sci Rep 2020 Dec 15;10(1):21976. [doi: 10.1038/s41598-020-78892-0] [Medline: 33319825]
133. Gill MB, Kutok JL, Fingeroth JD. Epstein-Barr virus thymidine kinase is a centrosomal resident precisely localized to the periphery of centrioles. J Virol 2007 Jun 15;81(12):6523-6535. [doi: 10.1128/JVI.00147-07] [Medline: 17428875]
134. Shumilov A, Tsai MH, Schlosser YT, et al. Epstein-Barr virus particles induce centrosome amplification and chromosomal instability. Nat Commun 2017 Feb 10;8:14257. [doi: 10.1038/ncomms14257] [Medline: 28186092]
135. Lingle WL, Lutz WH, Ingle JN, Maihle NJ, Salisbury JL. Centrosome hypertrophy in human breast tumors: implications for genomic stability and cell polarity. Proc Natl Acad Sci U S A 1998 Mar 17;95(6):2950-2955. [doi: 10.1073/pnas.95.6.2950] [Medline: 9501196]
136. Buschle A, Mrozek-Gorska P, Cernilogar FM, et al. Epstein-Barr virus inactivates the transcriptome and disrupts the chromatin architecture of its host cell in the first phase of lytic reactivation. Nucleic Acids Res 2021 Apr 6;49(6):3217-3241. [doi: 10.1093/nar/gkab099] [Medline: 33675667]

Abbreviations

BL: Burkitt lymphoma
BLAST: Basic Local Alignment Search Tool
bp: base pairs
COSMIC: Catalog of Somatic Mutations in Cancer
CRISPR: clustered regularly interspaced short palindromic repeats
DLBCL: diffuse large B-cell lymphoma
EBNA1: Epstein-Barr virus nuclear antigen 1
EBV: Epstein-Barr virus
FA: Fanconi anemia
GC: gastric cancer
HER2: human epidermal growth factor receptor 2
HERV: human endogenous retrovirus
HIV1: human immunodeficiency virus type 1
HLA: human leukocyte antigen
HPV: human papillomavirus
JAK: Janus kinase
MHC: major histocompatibility complex
NF-κB: nuclear factor–κB
NPC: nasopharyngeal cancer
piRNA: Piwi-interacting RNA
SWI-SNF : switch/sucrose non-fermentable
TIL: tumor-infiltrating lymphocyte

Edited by A Schwartz; submitted 10.07.23; peer-reviewed by Anonymous, Anonymous; revised version received 19.11.24; accepted 20.11.24; published 29.01.25.

Please cite as:
Friedenson B
Identifying Safeguards Disabled by Epstein-Barr Virus Infections in Genomes From Patients With Breast Cancer: Chromosomal Bioinformatics Analysis
JMIRx Med 2025;6:e50712
URL: https://xmed.jmir.org/2025/1/e50712
doi:10.2196/50712

Improved Alzheimer Disease Diagnosis With a Machine Learning Approach and Neuroimaging: Case Study Development

Lilia Lazli, PhD

Department of Computer and Software Engineering, Polytechnique Montréal, University of Montreal, 2500 Chem de Polytechnique, Montreal, QC, Canada

Corresponding Author:
Lilia Lazli, PhD
Department of Computer and Software Engineering, Polytechnique Montréal, University of Montreal, 2500 Chem de Polytechnique, Montreal, QC, Canada

Related Articles:

Companion article: https://arxiv.org/abs/2405.09553v1

Companion article: https://med.jmirx.org/2025/1/e73768

Companion article: https://med.jmirx.org/2025/1/e73454

Companion article: https://med.jmirx.org/2025/1/e73130

Companion article: https://med.jmirx.org/2025/1/e72821

Abstract

Background: Alzheimer disease (AD) is a severe neurological brain disorder. While not curable, earlier detection can help improve symptoms substantially. Machine learning (ML) models are popular and well suited for medical image processing tasks such as computer-aided diagnosis. These techniques can improve the process for an accurate diagnosis of AD.

Objective: In this paper, a complete computer-aided diagnosis system for the diagnosis of AD has been presented. We investigate the performance of some of the most used ML techniques for AD detection and classification using neuroimages from the Open Access Series of Imaging Studies (OASIS) and Alzheimer's Disease Neuroimaging Initiative (ADNI) datasets.

Methods: The system uses artificial neural networks (ANNs) and support vector machines (SVMs) as classifiers, and dimensionality reduction techniques as feature extractors. To retrieve features from the neuroimages, we used principal component analysis (PCA), linear discriminant analysis, and t-distributed stochastic neighbor embedding. These features are fed into feedforward neural networks (FFNNs) and SVM-based ML classifiers. Furthermore, we applied the vision transformer (ViT)–based ANNs in conjunction with data augmentation to distinguish patients with AD from healthy controls.

Results: Experiments were performed on magnetic resonance imaging and positron emission tomography scans. The OASIS dataset included a total of 300 patients, while the ADNI dataset included 231 patients. For OASIS, 90 (30%) patients were healthy and 210 (70%) were severely impaired by AD. Likewise for the ADNI database, a total of 149 (64.5%) patients with AD were detected and 82 (35.5%) patients were used as healthy controls. An important difference was established between healthy patients and patients with AD (P=.02). We examined the effectiveness of the three feature extractors and classifiers using 5-fold cross-validation and confusion matrix–based standard classification metrics, namely, accuracy, sensitivity, specificity, precision, F_1-score, and area under the receiver operating characteristic curve (AUROC). Compared with the state-of-the-art performing methods, the success rate was satisfactory for all the created ML models, but SVM and FFNN performed best with the PCA extractor, while the ViT classifier performed best with more data. The data augmentation/ViT approach worked better overall, achieving accuracies of 93.2% (sensitivity=87.2, specificity=90.5, precision=87.6, F_1-score=88.7, and AUROC=92) for OASIS and 90.4% (sensitivity=85.4, specificity=88.6, precision=86.9, F_1-score=88, and AUROC=90) for ADNI.

Conclusions: Effective ML models using neuroimaging data could help physicians working on AD diagnosis and will assist them in prescribing timely treatment to patients with AD. Good results were obtained on the OASIS and ADNI datasets with all the proposed classifiers, namely, SVM, FFNN, and ViTs. However, the results show that the ViT model is much better at predicting AD than the other models when a sufficient amount of data are available to perform the training. This highlights that the data augmentation process could impact the overall performance of the ViT model.

(JMIRx Med 2025;6:e60866) doi:10.2196/60866

KEYWORDS

Alzheimer disease; computer-aided diagnosis system; machine learning; principal component analysis; linear discriminant analysis; t-distributed stochastic neighbor embedding; feedforward neural network; vision transformer architecture; support vector machines; magnetic resonance imaging; positron emission tomography imaging; Open Access Series of Imaging Studies; Alzheimer's Disease Neuroimaging Initiative; OASIS; ADNI

Introduction

Alzheimer disease (AD) is a progressive degenerative brain disorder that gradually destroys memory, reason, judgment, language, and ultimately the ability to perform even the simplest of tasks [1]. An automated AD classification system is crucial for the early detection of disease. This computer-aided diagnosis (CAD) system can help expert clinicians prescribe the proper treatment and prevent brain tissue damage [1].

In the last decades, researchers have developed several CAD systems [1-5]. Rule-based expert systems were developed from the 1970s to the 1990s and supervised models from the 1990s [1]. Moreover, several approaches have been proposed in the literature aiming at providing an automatic tool that guides the clinician in the AD diagnosis process [1,5-7]. We can categorize these approaches into two types: univariate approaches, like statistical parametric mapping (SPM), and multivariate approaches, like the voxels-as-features (VAF) approach.

Due to advances in computing power, machine learning (ML) has encompassed many health care sectors and has shown results with organ and substructure segmentation as well as disease classifications in areas of pathology, brain, breast, bone, retina, etc. Open-access datasets on AD have led to the development of CAD systems that use ML to help scientists and medical staff make early diagnoses. These systems will ultimately help speed up the treatment of patients with AD. To make predictions, scientists have adopted various ML-based classifiers, including support vector machines (SVMs) [8,9], hidden Markov models [10,11], *k*-nearest neighbors classifier [12,13], discriminant analysis [14,15], random forest [16,17], decision trees [18], naive Bayes classifier [19,20], and artificial neural networks (ANNs) [21,22].

Despite the efforts of researchers, there have been few works on AD detection using ML models that have had significant performance, and the development of an automated AD classification model remains a rather challenging task. Within this framework of distinguishing between healthy controls (HCs) and people with AD, the main contributions of this paper can be summarized as follows.

- We developed a CAD system using the best-supervised learning classifiers, such as SVMs [8,9], feedforward neural networks (FFNNs) [23], and transformer neural networks, especially the vision transformer (ViT) architecture [24], which is becoming more popular in the field of computer vision due to its effectiveness.
- We designed these models to analyze the two neuroimages commonly used in AD diagnosis, namely, structural magnetic resonance imaging (sMRI) and fluorodeoxyglucose (FDG)–positron emission tomography (PET) as these modalities are the preeminent sources of information in the CAD process.
- The multimodal CAD system uses principal component analysis (PCA) [25] in conjunction with SVM and FFNN, training them on the PCA features extracted from the neurological images.
- The most challenging datasets, namely the Open Access Series of Imaging Studies (OASIS) [26] and Alzheimer's Disease Neuroimaging Initiative (ADNI) [27] datasets, underwent rigorous tests using various experimental settings. These experiments validated the effectiveness of the chosen models, showcasing their superiority over state-of-the-art approaches in terms of accuracy, sensitivity, specificity, precision, F_1-score, and area under the receiver operating characteristic curve (AUROC).

Methods

Participants

Sometimes we found signs of AD in the brain data of healthy and older patients, so considerable experience and knowledge were essential to distinguish the AD data from the HC patients' data. In this context, we have experimented the performance of the proposed CAD system on the OASIS [26] and ADNI [27] datasets.

OASIS Dataset

The OASIS dataset [26] was prepared by Dr Randy Buckner from the Howard Hughes Medical Institute at Harvard University, the Neuroinformatics Research Group at Washington University School of Medicine, and the Biomedical Informatics Research Network. OASIS is a longitudinal multimodal neuroimaging, clinical, cognitive, and biomarker dataset for normal aging and AD. We selected the patients with and without dementia from a larger database and obtained them from the longitudinal pool of the Washington University Alzheimer Disease Research Center. The experiment used a dataset that included 90 cognitively normal patients and 210 individuals with AD. The AD group included very mild, mild, moderate, and severe dementia.

ADNI Dataset

The ADNI dataset [27], which is the most commonly used in machine learning tasks, is an association of medical centers and universities located in the United States and Canada. ADNI is funded by the National Institute on Aging and the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie; Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc; Cogstate; Eisai

Co., Ltd; Elan Pharmaceuticals, Inc; Eli Lilly and Company; EUROIMMUN; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc; Fujirebio; GE HealthCare; IXICO plc; Janssen Alzheimer Immunotherapy Research & Development, LLC; Johnson & Johnson Pharmaceutical Research & Development, LLC; Lumosity; Lundbeck; Merck & Co., Inc; Meso Scale Diagnostics LLC; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health. The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuroimaging at the University of Southern California.

The main aim of ADNI is to provide open-source datasets to discover biomarkers and identify and track the progression of AD accurately. It developed to become an ideal source of longitudinal multisite PET and magnetic resonance imaging (MRI) images of patients with AD and older control patients (HC). The datasets were formed to make the detection system powerful by providing baseline information regarding changes in brain structure and metabolism, as well as clinical, cognitive, and biochemical data. The ADNI cohort used in our study included 82 cognitively normal patients and 149 patients with AD. The AD group included patients with mild cognitive impairment and those with confirmed AD.

Ethical Considerations

This work used two datasets (ADNI and OASIS), which are available in the public domain. For the benchmark ADNI dataset, the terms of use are declared on their website [28]. All patients in the ADNI database provided written informed consent, which was approved by the institutional review board of each participating institution. Patients were informed that their information would be kept confidential and their data would be anonymous and would be part of scientific publications.

According to local legislation and institutional requirements, the study of human participants using the OASIS dataset does not require ethical review and approval [26]. Written informed consent from the patients' legal guardians or next of kin was not required to participate in this study in accordance with national legislation and institutional requirements [26]. The data used for the analysis has been deidentified and made public.

Data Preparation

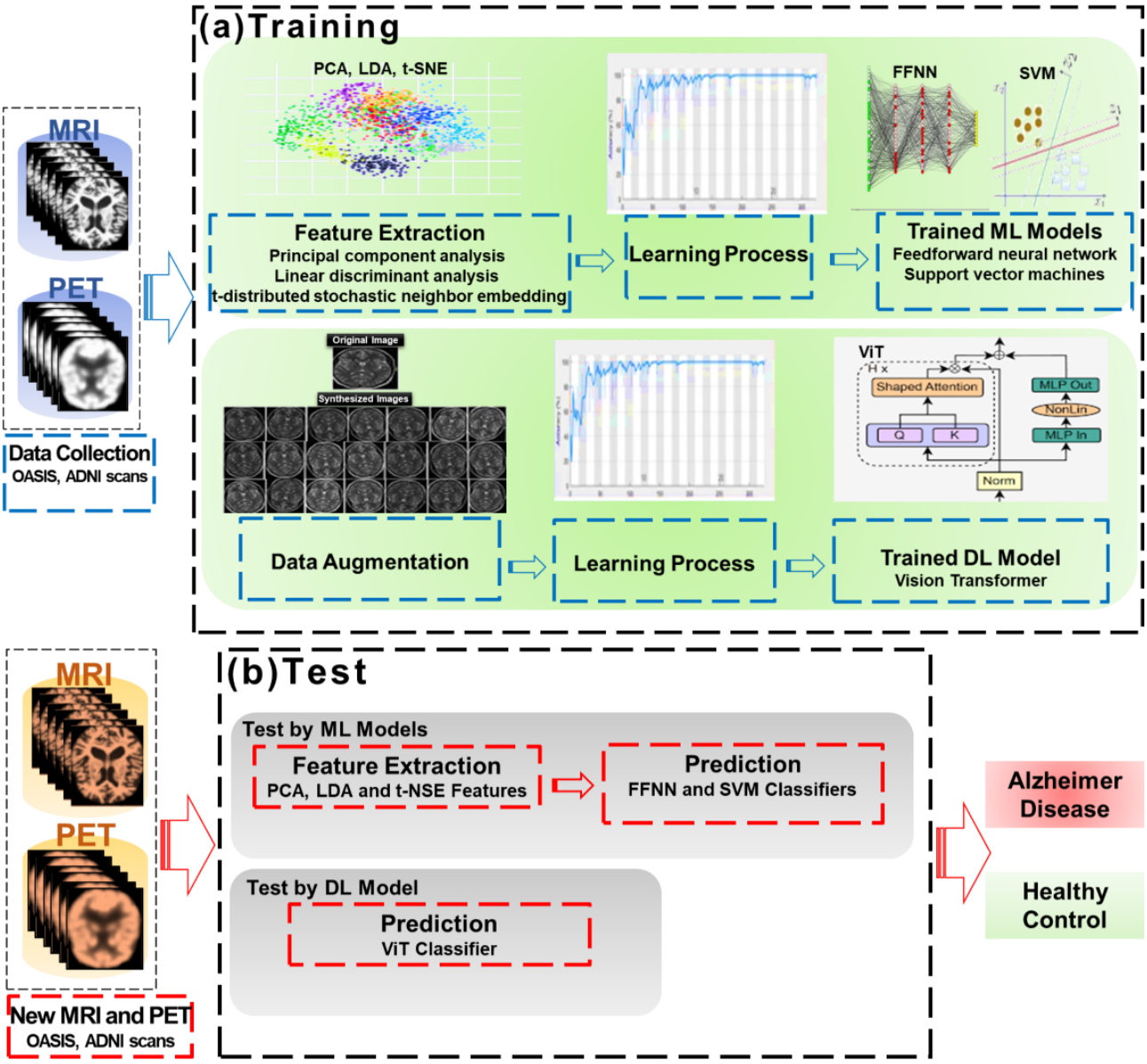
We performed the following steps on the OASIS and ADNI neuroimages: normalization, resizing, removing nonbrain slices, selecting slices with the most information, and converting 3D images into 2D slices. First, the damaged original files containing the images were removed. We selected a larger number of central slices to aid the CAD system in accurately classifying AD. We used an SPM tool (SPM8 [29]), which is a major update to SPM software, originally developed by Karl Friston, to partially correct spatial intensity inhomogeneities. This software normalized all the images using a general affine model with 12 parameters. The origin of the raw sMRI scans was set manually to anterior commissure before manually registering them with SPM's canonical T1 template image. We applied the nonparametric nonuniform intensity normalization (N3) technique to solve the tissue intensity nonuniformity problem [30]. Then the hybrid median filter was used to remove impulse noise while preserving edges.

ML Approaches

Overview

A generic automated AD detection and classification framework is summarized in Figure 1. ML classifiers aim to predict the class of the input data (images of patients with AD or healthy patients) by looking at a number of learning examples. The process begins with the preprocessing of sMRI and FDG-PET images to keep only relevant data. Then each image is represented by grayscale features and is collapsed into a new feature space by applying PCA-based feature extraction to pick the optimal features. After that, to classify the patients, these selected features are fed to the supervised learner. In this work, SVMs and FFNNs are learned on the PCA features extracted from the neuroimages. While for ViT, we applied the data augmentation strategy [31], since the training of this network required more data compared to the other two classifiers. For PCA, a performance comparison was made with similar techniques, t-distributed stochastic neighbor embedding (t-SNE) [32] and linear discriminant analysis (LDA) [14].

Figure 1. Block diagram of a generic Alzheimer disease computer-aided diagnosis system. ADNI: Alzheimer's Disease Neuroimaging Initiative; DL: deep learning; FFNN: feedforward neural network; LDA: linear discriminant analysis; ML: machine learning; MRI: magnetic resonance imaging; OASIS: Open Access Series of Imaging Studies; PCA: principal component analysis; PET: positron emission tomography; SVM: support vector machine; t-SNE: t-distributed stochastic neighbor embedding; ViT: vision transformer.



Below is a summary description of the four approaches proposed for our CAD system, and more details on the mathematical background of these approaches can be found in Multimedia Appendix 1 for PCA, Multimedia Appendix 2 for SVM, Multimedia Appendix 3 for FFNN, and Multimedia Appendix 4 for ViT.

Principal Component Analysis

PCA is a linear dimensionality reduction method used widely in data preprocessing and exploratory analysis. Different image classification purposes have successfully used PCA because its method is nonparametric and easy to apply, and helps extract useful information from confusing datasets [25].

In this study, we used this technique to extract useful features for classifiers. PCA allows the production of new variables that represent linear combinations of the original variables. Using linear algebra and matrix operations, a transformation is performed from the original dataset to a new coordinate system structured by the principal components. The analysis of this linear transformation is obtained thanks to the eigenvectors and the eigenvalues of the covariance matrix. The PCA steps are summarized as follows: (1) standardize the range of continuous initial variables, (2) find correlations by computing the covariance matrix, (3) find the eigenvectors and eigenvalues of the covariance matrix, (4) choose the principal components, and (5) change the data to the new coordinate system. More details about the PCA computation process with mathematical formulas are explained in Multimedia Appendix 1.

Support Vector Machines

We used SVMs as classifiers for the classification of independent and identically distributed data [23]. These machines are widely used as supervised max-margin models, along with associated learning algorithms that analyze data. To distinguish two classes, the principle of SVMs is to seek the

optimal hyperplane that allows for maximizing the margin between the closest data points of the opposite classes.

The SVM algorithm for linear classification is widely used in ML. However, in this study, we used SVMs to perform nonlinear classification due to the data's nonlinear separability. We achieved this by applying a kernel function to represent the data as a set of pairwise similarity comparisons between the original data points.

This function transforms the original data points into coordinates in a higher-dimensional feature space, thereby facilitating linear separation. Multimedia Appendix 2 provides further details about the SVM computation process, including mathematical formulas.

Feedforward Neural Network

Biological nervous systems, such as the brain, inspire the information-processing paradigm of FFNN, which is one of the two main types of ANNs [23]. The distinctive feature of this network is the unidirectional flow of information, meaning that the information flow in the model is only in one direction—forward—without any loops or cycles. Information flows from the input nodes through the hidden nodes and to the output nodes.

This network is static and memoryless. Given a data input, FFNN provides a single set of output values instead of a sequence of values. Furthermore, the response produced for an input is independent of the previous state of the network. FFNN automatically learns from examples and uses a backpropagation learning algorithm for determining weights. More details about the FFNN computation process with mathematical foundations are explained in Multimedia Appendix 3.

Transformers

Transformers, which dominate natural language processing, have acquired a reputation in computer vision owing to their positive results in many applications such as semantic segmentation, object detection, and image classification. Transformer architecture entirely relies on an attention mechanism to produce global dependencies between input and output, avoiding recurrence. Self-attention assesses the sequence representation by connecting various positions within a single sequence.

In this work, we applied a ViT architecture [24] to neuroimages with very little adjustment, demonstrating better performance in numerous computer-vision tasks. ViT uses a multiheaded self-attention mechanism to catch and learn long-range dependencies between distant positions by averaging attention-weighted positions. This promotes the network's focus on all of the data of the input sequence. This characteristic encourages us to use ViT for our brain imaging study owing to its capacity to precisely catch interdependencies between spreaded brain regions. More details about the ViT computation process with mathematical foundations are explained in Multimedia Appendix 4.

Nevertheless, the learning dataset is too small, involving substantial data to learn a ViT from scratch. In this regard, we used data augmentation to expand the size of the input data by creating additional data from the original input data. To create new images, we performed some geometric transformations. The visual transformation primarily focuses on translating, flipping random images horizontally, rotating them at 15 angles without cropping, and rescaling the input data to the range of [0, 1].

Statistical Analysis

We have carried out the performance assessment and the comparison of the classifiers using typical confusion matrix–based evaluation metrics. The confusion matrix has the elements of true positive (TP), false positive (FP), false negative (FN), and true negative (TN). Each column of the matrix indicates an instance of the predicted class, and each row contains a true (correct or actual) class. The following are the metrics used to evaluate the performance of the CAD system.

Sensitivity—also known as recall—is used for calculating the classifier's ability to correctly predict Alzheimer instances (AD class). On the other hand, the classifier uses specificity to accurately predict all non-Alzheimer instances (HC class) across all inputs.

A classifier should have high sensitivity and specificity. Therefore, the accuracy metric, which calculates the number of correctly classified instances relative to the total number of instances, is the average of these two measures. The precision metric measures the classifier's ability to quantify the number of TPs of the AD class that receive a correct label in classification.

The combined harmonic mean of both sensitivity and precision gives the F_1-score, which takes a value between 0 and 1. The receiver operating characteristic curve, a method for visualizing a classifier's ability to diagnose or predict correctly, clearly illustrates the trade-off that arises between the sensitivity and specificity metrics. At various thresholds, the receiver operating characteristic curve plots the TP rate or sensitivity against the FP rate (1 – specificity).

We aim to determine the degree of separability, or the ability to correctly predict class, using the AUROC. The higher the AUROC, the better; 1 would be perfect, and 0.5 would be random. Accuracy, sensitivity, specificity, precision, F_1-score, and AUROC are the six main metrics used to assess the efficacy of each classifier. The following are the mathematical formulas for the first five metrics.

(1)Accuracy=TP+TNTP+FP+FN+TP

(2)Sensitivity=TPTP+FN

(3)Specificity=TNTN+FP

(4)Precision=TPTP+FP

(5)F1−score=2Precision×SensitivityPrecision+Sensitivity

Results

We experimented the performance of the proposed CAD system on patients' images from the OASIS [26] and ADNI [27] datasets. These datasets contain sMRI and FDG-PET scans along with information about the patients' demographics and clinical assessments. There are 300 patients for OASIS and 231

patients for ADNI whose age was between 18 and 96 years, and each patient had 3 or 4 accessible PET and T1-weighted MRI scans. Tables 1 and 2 provide more details on the demographic and clinical characteristics of participants.

Table . The demographic information (gender, race, class, right-handed) of participants.

Variable	OASIS[a] patients (n=300), n (%)	ADNI[b] patients (n=231), n (%)
Gender		
Women	80 (26.7)	99 (42.9)
Men	220 (73.3)	132 (57.1)
Race		
Caucasian	174 (58.0)	159 (68.8)
African-American	122 (40.7)	70 (30.3)
Asian	4 (1.3)	2 (0.9)
Class		
Alzheimer	210 (70.0)	149 (64.5)
Healthy	90 (30.0)	82 (35.5)
Right-handed		
Women	77 (96.3)	93 (93.9)
Men	219 (99.5)	130 (98.5)

[a]OASIS: Open Access Series of Imaging Studies.

[b]ADNI: Alzheimer's Disease Neuroimaging Initiative.

Table . The demographic characteristics and clinical assessment data in terms of age, education, mini-mental state examination, and Alzheimer's Disease Assessment Scale–Cognitive subscale.

Variable	OASIS[d] patients, mean (SD; range)	ADNI[e] patients, mean (SD; range)
Age (years)		
Women	67.78 (43.2 - 95.6)	75.3 (5.2)
Men	70.17 (42.5 - 91.7)	75.4 (7.1)
Education		
Women	14.3 (1.6; 9-18)	15.6 (3.2)
Men	15.2 (2.7; 8-23)	14.9 (3.4)
Mini-mental state examination[f]		
Baseline (women)	25.4 (0.4; 22-26)	29.0 (1.2; 19-26)
2 years (women)	__[g]	29.0 (1.3)
Baseline (men)	23.8 (1.9; 25-29)	23.8 (1.9; 25–29)
2 years (men)	19.3 (5.6)	29.0 (1.2; 19-26)
Alzheimer's Disease Assessment Scale–Cognitive subscale[h]		
Baseline (women)	—	7.3 (3.3)
2 years (women)	—	6.3 (3.5)
Baseline (men)	—	7.3 (3.3)
2 years (men)	—	27.3 (11.7)

[d]OASIS: Open Access Series of Imaging Studies.

[e]ADNI: Alzheimer's Disease Neuroimaging Initiative.

[f]The mini-mental state examination has a possible score range of 0-30.

[g]Not available.

[h]The Alzheimer's Disease Assessment Scale–Cognitive subscale has a possible score range of 0-30.

We used a clinical dementia rating scale to control the dementia status of the dataset; a score of 0 on the scale indicates a normal cognitive level, while a score greater than 0 determines the presence of AD. In this context, we divided the images into 210 (70%) patients with AD and 90 (30%) HCs for the OASIS dataset and 149 (64.5%) patients with AD and 82 (35.5%) HCs for the ADNI dataset. The majority of the samples were identified as men, specifically 220 (73%) for OASIS and 132 (57%) for ADNI, while the majority of the samples were Caucasian, specifically 174 (58%) for OASIS and 159 (69%) for ADNI.

After the preprocessing steps, each slice of sMRI includes 256 × 256 × 176 voxels covering the entire region of the brain with the following parameters: voxel size is $2 \times 2 \times 2$ mm^3 for ADNI and $2 \times 3.1 \times 2$ mm^3 for OASIS, isotropic resolution is 1.0 mm, time of repetition is 5050 milliseconds, and time of echo is 10 milliseconds. All slices of reconstructed PET images are resampled to contain 256 × 256 × 207 voxels with a voxel size of $1.2 \times 1.2 \times 1.2$ mm^3.

The appropriate hyperparameter values for the classifiers were chosen by reviewing prior state-of-the-art work and after doing empirical testing and exploratory analyses. Some of the hyperparameters used in the experiment are presented in Table 3.

Table . The hyperparameter tuning and classifiers configuration used in the experiment.

Hyperparameter	Search range
Support vector machine	
Multiclass method	One-vs-one (one-vs-all, one-vs-one)
Penality parameter of error	0.001 (0.0001, 0.001, 0.01, 0.1)
Box constraint level	1 (0.001 - 1000)
Kernel function	Gaussian (Gaussian, linear, quadratic, cubic)
Kernel scale	2.8
Iteration	30
Standardize data	True
Feedforward neural network	
Number of fully connected layers	1
First layer size	100
Activation	Hyperbolic tangent sigmoid
Learning function	Gradient descent with momentum weight and bias
Iteration limit	1000
Regularizarion strength (λ)	0
Update of weight and bias	Levenberg-Marquardt optimization
Standardize data	True
Vision transformer	
Layers	12
Hidden size D	768
Multilayer perceptron size	3072
Heads	12
Parameters	86 million
Path resolution	16 × 16

For training and testing, 5-fold cross-validation was achieved on each dataset. For each fold, 70% of the data was used for training, 10% for validation, and 20% for testing the effectiveness of each classifier. We conducted experiments on SVM and FFNN using four dimensionality reduction techniques (VAF, LDA, t-SNE, and PCA), as well as on the ViT classifier, without and with data augmentation. During the training process, SVM and FFNN achieved the best results with PCA for the validation data, while the ViT classifier achieved the best results with increased data.

For the test data, we obtained for the OASIS dataset an accuracy of 91.9% (prediction speed ~2000 observations/second, training time 1.5703 seconds) for SVM, 88.2% (prediction speed ~6000 observations/second, training time 7.7715 seconds) for FFNN, and 93.2% (prediction speed ~7000 observations/second, training time 102.3529 seconds) for ViT. The same result was seen for the ADNI data, with an accuracy of 88.6% for SVM

(prediction speed ~1300 observations/second, training time 1.4280 seconds), 80.9% for FFNN (prediction speed ~5300 observations/second, training time 8.2319 seconds), and 90.4% for ViT (prediction speed ~7200 observations/second, training time 129.4531 seconds). Tables 4 and 5 provide further details about the top classification results achieved with the proposed ML classifiers for the OASIS and ADNI datasets, respectively, based on six metrics.

Table . Five-fold cross-validation performance for the Open Access Series of Imaging Studies test data in terms of accuracy, sensitivity, specificity, precision, F_1-score, and area under the receiver operating characteristic curve (AUROC).

Classifier	Accuracy (%)	Sensitivity (%)	Specificity (%)	Precision (%)	F_1-score (%)	AUROC (%)
Support vector machine						
VAF[a]	66.3	61.3	62.1	65.1	52.4	60
LDA[b]	75.6	70.1	69	70.6	68.7	72
t-SNE[c]	80.2	74.5	72.4	71.4	70.1	73
PCA[d]	*91.9* [e]	*86.4*	*90.6*	*87.2*	*89*	*90*
Feedforward neural network						
VAF	62.4	54.1	57.2	51.6	53.4	51
LDA	70.5	66.4	71.4	68.9	72.5	66
t-SNE	72.6	71.3	70.2	69.4	72.8	73
PCA	88.2	85.4	84.6	86.2	83.7	82
Vision transformer						
Without data augmentation	60.8	53.1	54.6	56.8	55.6	61
With data augmentation	*93.2*	*87.2*	*90.5*	*87.6*	*88.7*	*92*

[a]VAF: voxels-as-features.

[b]LDA: linear discriminant analysis.

[c]t-SNE: t-distributed stochastic neighbor embedding.

[d]PCA: principal component analysis.

[e]Italics indicate the best achieved results.

Table . Five-fold cross-validation performance for Alzheimer's Disease Neuroimaging Initiative test data in terms of accuracy, sensitivity, specificity, precision, F_1-score, and area under the receiver operating characteristic curve (AUROC).

Classifier	Accuracy (%)	Sensitivity (%)	Specificity (%)	Precision (%)	F_1-score (%)	AUROC (%)
Support vector machine						
VAF[a]	42.8	59.2	60.4	63.2	50.1	58
LDA[b]	72.1	68.4	67.2	68.4	66.2	70
t-SNE[c]	79.3	71.1	70.1	69.2	68.3	71
PCA[d]	*88.6*[e]	*84.1*	*88.4*	*85.1*	*87.4*	*88*
Feedforward neural network						
VAF	60.9	51.3	56.4	49.1	51	48
LDA	69.1	62.3	70	65.4	70.1	63
t-SNE	70.4	68.1	68.4	67.1	70.4	70
PCA	80.9	84.1	82.3	84.3	81.4	80
Vision transformer						
Without data augmentation	59.3	50.2	51.1	54.4	53.4	57
With data augmentation	*90.4*	*85.4*	*88.6*	*86.9*	*88*	*90*

[a]VAF: voxels-as-features.

[b]LDA: linear discriminant analysis.

[c]t-SNE: t-distributed stochastic neighbor embedding.

[d]PCA: principal component analysis.

[e]Italics indicate the best achieved results.

Discussion

Main Findings

The main finding is that the development of diagnostic tools applying the ML approach in conjunction with neuroimaging data could substantially help in automating the classification and prediction of AD.

In this context, this study proposed a complete CAD system to successfully classify patients with AD and discriminate them from HC patients. The purpose was to examine the association between SVM, FFNN, and ViT ML classifiers; PCA, LDA, and t-SNE dimensionality reduction techniques; and sMRI and FDG-PET neuroimaging modalities to detect early signs of AD. Furthermore, we aimed to clarify the impact of some data preprocessing strategies, such as noise reduction and data augmentation, on improving the performance of classifiers.

With regard to the sMRI and FDG-PET modalities, they can provide large amounts of information; nevertheless, interpreting all image content is challenging for physicians. The experimental analysis demonstrates that combining these neuroimaging modalities with selected ML classifiers enhances their performance, enabling doctors to provide precise diagnosis and timely patient care. This confirms the theory regarding the benefits of these two modalities. Since sMRI provides high-resolution images of brain anatomical structures, which confirm structural change in the brain, it shows shrinkage of brain tissue and abnormalities, while FDG-PET shows the functionality of the brain.

Regarding the selected dimensional reduction techniques, all of the chosen dimensional reduction techniques performed well as feature extractors when combined with the SVM and FFNN classifiers, but a comparative analysis of the three techniques reveals that PCA outperforms LDA and t-SNE. However, it is important to clarify certain findings: PCA allows the identification of the most significant variables in the data due to its potential to generate new variables, which represent linear combinations of the original variables. Moreover, t-SNE differs from PCA by preserving only small pairwise distances or local similarities, while PCA aims to preserve large pairwise distances to maximize variance. Unlike PCA, LDA is a supervised technique that maximizes class separability in the reduced dimensionality space, thereby retaining the most discriminative features.

Preliminary results from evaluating the complete CAD system using the three classifiers prove that the system is more effective in separating AD and HC classes. The results provided by all the experiments carried out reveal an increase in sensitivity and, consequently, the final accuracy obtained by the basic VAF-SVM model (66.3% for OASIS and 42.8% for ADNI). We compared the performance of the SVM, FFNN, and ViT models using confusion matrix–based metrics.

All models performed well, providing acceptable performance for both databases. Data augmentation/ViT outperformed other

models, with accuracies of 93.2% for OASIS and 90.4% for ADNI (see Tables 4 and 5 for more details on results obtained from all models tested on both databases). The second best classifier is PCA/SVM, achieving an accuracy decrease of 1.3% for OASIS and 1.8% for ADNI, compared to the rates obtained by ViT, resulting in overall accuracy rates of 91.9% and 88.6% for OASIS and ADNI, respectively. Therefore, the data augmentation process and the PCA dimensionality reduction method have the potential to impact the overall performance of the ViT and SVM models, respectively.

Moreover, compared to the performance using a single MRI modality, all models performed well using a multimodal MRI/PET environment. The best results with MRI were also obtained with ViT and SVM classifiers. Accuracies of 83.9% for the OASIS dataset and 81.2% for ADNI were obtained using the data augmentation/ViT approach. PCA/SVM achieved accuracies of 82.4% for the OASIS and 80.6% for the ADNI datasets. This draws attention to the potential of integrating multiple modalities to increase the performance of the CAD system.

Comparison With Prior Work

To verify the convergence of the proposed CAD system, we compared the results obtained with some relevant state-of-the-art ML models. The experimental results show that our models, particularly SVM and ViT, have good performance on both the OASIS and ADNI datasets and achieved better or comparable accuracy to most existing methods in the literature. For the OASIS dataset, the PCA/SVM method had a 91.9% accuracy and the ViT model with data augmentation had a 93.2% accuracy. Nanni et al [33], Khan and Zubair [16], Sethi et al [2], Basheer et al [34], Saratxaga et al [35], and Liu et al [36] got 90.2%, 86.8%, 86.2%, 92.3%, 93%, and 82.6% accuracy, respectively.

The same finding was obtained for the ADNI dataset, where we achieved an accuracy of 88.6% using the PCA/SVM approach and 90.4% using the ViT model by increasing the data. In contrast, the accuracy achieved by Rallabandi et al [37], Jo et al [4], Jo et al [3], Liu et al [36], and Shojaei et al [38] was 75%, 75.02%, 80.8%, 90%, and 87%, respectively. Table 6 compares our best results obtained with the prior state-of-the-art models discussed.

Table . Comparative study of performance with state-of-the-art machine learning models using the Open Access Series of Imaging Studies (OASIS) and Alzheimer's Disease Neuroimaging Initiative (ADNI) datasets.

Study	Approach	Dataset	Accuracy	Sensitivity	F_1-score	AUROC[a]
Liu et al [36]	Monte Carlo sampling/ResNet50-CNNs[b]/ensemble classifier	OASIS	82.6	74.3	—[c]	—
Saratxaga et al [35]	ResNet18-based CNNs	OASIS	93	—	—	—
Basheer et al [34]	PCA[d]/ CapsNet-based CNNs	OASIS	92.3	82.3	—	—
Nanni et al [33]	Ensemble of 5 transfer learning models	OASIS	90.2	—	—	—
Khan and Zubair [16]	Chi-square statistical test/RF[e]	OASIS	86.8	80	86.4	87.2
Sethi et al [2]	CNNs/ SVM[f]	OASIS	86.2	—	—	—
Our study	PCA/SVM	OASIS	91.9	86.4	89	90
Our study	Data augmentation/ViT[g]	OASIS	*93.2*[h]	*87.2*	*88.7*	*92*
Shojaei et al [38]	Genetic algorithm/3D-CNNs	ADNI	87	—	—	—
Liu et al [36]	Monte Carlo sampling/ResNet50-CNNs/ensemble classifier	ADNI	90	83.5	—	—
Rallabandi et al [37]	FreeSurfer/SVM	ADNI	75	75	72	76
Jo et al [4]	Sliding Window Association Test/CNNs	ADNI	75	—	—	82
Jo et al [3]	Weighted gene co-expression network analysis/RF	ADNI	80.8	—	—	80.8
Our study	PCA/SVM	ADNI	88.6	84.1	87.4	88
Our study	Data augmentation/ViT	ADNI	*90.4*	*85.4*	*88*	*90*

[a]AUROC: area under the receiver operating characteristic curve.

[b]CNN: convolutional neural network.

[c]Not available.

[d]PCA: principal component analysis.

[e]RF: random forest.

[f]SVM: support vector machine.

[g]ViT: vision transformer.

[h]Italics indicate the best achieved results.

Limitations and Future Directions

There are several improvements possible for the proposed CAD system. We aim to enhance the system's performance by collaborating with more extensive AD datasets and implementing various types of ANN and ML-based classifiers.

The PCA used for feature extraction looks for the principal axis direction, which is used to effectively represent the common features of similar samples. This is very effective for representing the common features of the same kind of data samples, but it is not suitable for distinguishing different sample classes. Therefore, to achieve the purpose of feature extraction, we need to combine PCA with other feature dimensionality reduction algorithms like uniform manifold approximation and projection.

Acknowledgments

This project was supported by “Fonds de recherche du Québec-Nature et Technologies -FRQNT” grants under awards B3X × 314498 and B3XR × 358107.

The author would like to thank “Fonds de recherche du Québec-Nature et Technologies -FRQNT” for the financial support offered to accomplish this project. Many thanks to the researchers and expert clinicians of the Open Access Series of Imaging Studies and Alzheimer’s Disease Neuroimaging Initiative (ADNI) datasets for developing the images used in the preparation of this work. Special thanks to the reviewers and proofreader (Joshua Dykas) of this work.

Data used in preparation of this article were obtained from the ADNI database. As such, the investigators within the ADNI contributed to the design and implementation of ADNI or provided data but did not participate in the analysis or writing of this report. A complete listing of ADNI investigators can be found at [39].

Data Availability

This study used two datasets, Open Access Series of Imaging Studies [26] and Alzheimer’s Disease Neuroimaging Initiative [27], which are available in the public domain. However, they are subject to restrictions because they were used under permissions for this study and are therefore not publicly available.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Principal component analysis.
[DOCX File, 18 KB - xmed_v6i1e60866_app1.docx]

Multimedia Appendix 2
Support vector machines.
[DOCX File, 16 KB - xmed_v6i1e60866_app2.docx]

Multimedia Appendix 3
Feedforward neural network.
[DOCX File, 18 KB - xmed_v6i1e60866_app3.docx]

Multimedia Appendix 4
Vision transformer.
[DOCX File, 18 KB - xmed_v6i1e60866_app4.docx]

References

1. Hu X, Sun Z, Nian Y, et al. Self-explainable graph neural network for Alzheimer disease and related dementias risk prediction: algorithm development and validation study. JMIR Aging 2024 Jul 8;7:e54748. [doi: 10.2196/54748] [Medline: 38976869]
2. Sethi M, Rani S, Singh A, Mazón JLV. A CAD system for Alzheimer’s disease classification using neuroimaging MRI 2D slices. Comput Math Methods Med 2022 Aug 9;2022:8680737. [doi: 10.1155/2022/8680737] [Medline: 35983528]
3. Jo T, Kim J, Bice P, et al. Circular-SWAT for deep learning based diagnostic classification of Alzheimer’s disease: application to metabolome data. EBioMedicine 2023 Nov;97:104820. [doi: 10.1016/j.ebiom.2023.104820] [Medline: 37806288]
4. Jo T, Nho K, Bice P, Saykin AJ, Alzheimer’s Disease Neuroimaging Initiative. Deep learning-based identification of genetic variants: application to Alzheimer’s disease classification. Brief Bioinform 2022 Mar 10;23(2):bbac022. [doi: 10.1093/bib/bbac022] [Medline: 35183061]
5. Lazli L, Boukadoum M, Mohamed OA. A survey on computer-aided diagnosis of brain disorders through MRI based on machine learning and data mining methodologies with an emphasis on Alzheimer disease diagnosis and the contribution of the multimodal fusion. Appl Sci (Basel) 2020;10(5):1894. [doi: 10.3390/app10051894]
6. Lazli L, Boukadoum M, Ait Mohamed O. Computer-aided diagnosis system of Alzheimer’s disease based on multimodal fusion: tissue quantification based on the hybrid fuzzy-genetic-possibilistic model and discriminative classification based on the SVDD model. Brain Sci 2019 Oct 22;9(10):289. [doi: 10.3390/brainsci9100289] [Medline: 31652635]

7. Groppell S, Soto-Ruiz KM, Flores B, et al. A rapid, mobile neurocognitive screening test to aid in identifying cognitive impairment and dementia (BrainCheck): cohort study. JMIR Aging 2019 Mar 21;2(1):e12615. [doi: 10.2196/12615] [Medline: 31518280]
8. Eke CS, Jammeh E, Li X, Carroll CB, Pearson S, Ifeachor EC. Early detection of Alzheimer's disease with blood plasma proteins using support vector machines. IEEE J Biomed Health Inform 2021 Jan;25(1):218-226. [doi: 10.1109/JBHI.2020.2984355] [Medline: 32340968]
9. Cai W, Chu C, Zhu X. Intelligent classification of Alzheimer's disease based on support vector machine. Presented at: 3rd International Conference on Applied Mathematics, Modelling and Intelligent Computing (CAMMIC 2023); Mar 24-26, 2023; Tangshan, China. [doi: 10.1117/12.2686137]
10. Kang K, Cai J, Song X, Zhu H. Bayesian hidden Markov models for delineating the pathology of Alzheimer's disease. Stat Methods Med Res 2019 Jul;28(7):2112-2124. [doi: 10.1177/0962280217748675] [Medline: 29278101]
11. Tahami Monfared AA, Fu S, Hummel N, et al. Estimating transition probabilities across the Alzheimer's disease continuum using a nationally representative real-world database in the United States. Neurol Ther 2023 Aug;12(4):1235-1255. [doi: 10.1007/s40120-023-00498-1] [Medline: 37256433]
12. Elgammal YM, Zahran MA, Abdelsalam MM. A new strategy for the early detection of alzheimer disease stages using multifractal geometry analysis based on k-nearest neighbor algorithm. Sci Rep 2022 Dec 26;12(1):22381. [doi: 10.1038/s41598-022-26958-6] [Medline: 36572791]
13. Lu D, Yue Y, Hu Z, Xu M, Tong Y, Ma H. Effective detection of Alzheimer's disease by optimizing fuzzy k-nearest neighbors based on salp swarm algorithm. Comput Biol Med 2023 Jun;159:106930. [doi: 10.1016/j.compbiomed.2023.106930] [Medline: 37087779]
14. Jha D, Alam S, Pyun JY, Lee KH, Kwon GR. Alzheimer's disease detection using extreme learning machine, complex dual tree wavelet principal coefficients and linear discriminant analysis. J Med Imaging Health Inform 2018 Jun 1;8(5):881-890. [doi: 10.1166/jmihi.2018.2381]
15. Lin W, Gao Q, Du M, Chen W, Tong T. Multiclass diagnosis of stages of Alzheimer's disease using linear discriminant analysis scoring for multimodal data. Comput Biol Med 2021 Jul;134:104478. [doi: 10.1016/j.compbiomed.2021.104478] [Medline: 34000523]
16. Khan A, Zubair S. An improved multi-modal based machine learning approach for the prognosis of Alzheimer's disease. J King Saud Univ Comput Inf Sci 2022 Jun;34(6):2688-2706. [doi: 10.1016/j.jksuci.2020.04.004]
17. Shastry KA, Sattar SA. Logistic random forest boosting technique for Alzheimer's diagnosis. Int J Inf Technol 2023;15(3):1719-1731. [doi: 10.1007/s41870-023-01187-w] [Medline: 37056794]
18. Costa A, Pais M, Loureiro J, et al. Decision tree-based classification as a support to diagnosis in the Alzheimer's disease continuum using cerebrospinal fluid biomarkers: insights from automated analysis. Braz J Psychiatry 2022 Aug 30;44(4):370-377. [doi: 10.47626/1516-4446-2021-2277] [Medline: 35739065]
19. Bhagya Shree SR, Sheshadri HS. Diagnosis of Alzheimer's disease using naive Bayesian classifier. Neural Comput Applic 2018 Jan;29(1):123-132. [doi: 10.1007/s00521-016-2416-3]
20. Chandra A, Roy S. On the detection of alzheimer's disease using naïve bayes classifier. Presented at: 2023 International Conference on Microwave, Optical, and Communication Engineering (ICMOCE); May 26-28, 2023; Bhubaneswar, India. [doi: 10.1109/ICMOCE57812.2023.10166516]
21. Amoroso N, Diacono D, Fanizzi A, et al. Deep learning reveals Alzheimer's disease onset in MCI subjects: results from an international challenge. J Neurosci Methods 2018 May 15;302:3-9. [doi: 10.1016/j.jneumeth.2017.12.011] [Medline: 29287745]
22. Lella E, Lombardi A, Amoroso N, et al. Machine learning and DWI brain communicability networks for Alzheimer's disease detection. Appl Sci (Basel) 2020;10(3):934. [doi: 10.3390/app10030934]
23. Kotsiantis SB, Zaharakis ID, Pintelas PE. Machine learning: a review of classification and combining techniques. Artif Intell Rev 2006 Nov 10;26(3):159-190. [doi: 10.1007/s10462-007-9052-3]
24. Boesch G. Vision transformers (ViT) in image recognition. viso.ai. 2023 Nov 25. URL: https://viso.ai/deep-learning/vision-transformer-vit/ [accessed 2025-03-20]
25. Jolliffe IT, Cadima J. Principal component analysis: a review and recent developments. Philos Trans A Math Phys Eng Sci 2016 Apr 13;374(2065):20150202. [doi: 10.1098/rsta.2015.0202] [Medline: 26953178]
26. Open Access Series of Imaging Studies (OASIS). WashU Sites. URL: https://sites.wustl.edu/oasisbrains/ [accessed 2021-11-16]
27. Alzheimer's Disease Neuroimaging Initiative. URL: http://www.adni-info.org/ [accessed 2021-11-16]
28. Terms of use. Alzheimer's Disease Neuroimaging Initiative. URL: http://adni.loni.usc.edu/terms-of-use/ [accessed 2021-11-16]
29. SPM8. Wellcome Centre for Human Neuroimaging. URL: https://www.fil.ion.ucl.ac.uk/spm/software/spm8/ [accessed 2022-10-20]
30. Sled JG. The MNI_N3 software package. McConnell Brain Imaging Centre. 2004 Mar 15. URL: http://www.bic.mni.mcgill.ca/software/N3/ [accessed 2022-10-23]

31. Garcea F, Serra A, Lamberti F, Morra L. Data augmentation for medical imaging: a systematic literature review. Comput Biol Med 2023 Jan;152:106391. [doi: 10.1016/j.compbiomed.2022.106391] [Medline: 36549032]
32. Maaten L, Hinton G. Viualizing data using t-SNE. J Mach Learn Res 2008;9:2579-2605 [FREE Full text]
33. Nanni L, Interlenghi M, Brahnam S, et al. Comparison of transfer learning and conventional machine learning applied to structural brain MRI for the early diagnosis and prognosis of Alzheimer's disease. Front Neurol 2020 Nov 5;11:576194. [doi: 10.3389/fneur.2020.576194] [Medline: 33250847]
34. Basheer S, Bhatia S, Sakri SB. Computational modeling of dementia prediction using deep neural network: analysis on OASIS dataset. IEEE Access 2021 Mar 17;9:42449-42462. [doi: 10.1109/ACCESS.2021.3066213]
35. Saratxaga CL, Moya I, Picón A, et al. MRI deep learning-based solution for Alzheimer's disease prediction. J Pers Med 2021 Sep 9;11(9):902. [doi: 10.3390/jpm11090902] [Medline: 34575679]
36. Liu C, Huang F, Qiu A, Alzheimer's Disease Neuroimaging Initiative. Monte Carlo ensemble neural network for the diagnosis of Alzheimer's disease. Neural Netw 2023 Feb;159:14-24. [doi: 10.1016/j.neunet.2022.10.032] [Medline: 36525914]
37. Rallabandi VPS, Tulpule K, Gattu M. Automatic classification of cognitively normal, mild cognitive impairment and Alzheimer's disease using structural MRI analysis. Inform Med Unlocked 2020;18:100305. [doi: 10.1016/j.imu.2020.100305]
38. Shojaei S, Saniee Abadeh M, Momeni Z. An evolutionary explainable deep learning approach for Alzheimer's MRI classification. Expert Syst Appl 2023 Jun 15;220:119709. [doi: 10.1016/j.eswa.2023.119709]
39. Acknowledgement list for ADNI publications. Alzheimer's Disease Neuroimaging Initiative. URL: https://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf [accessed 2025-04-09]

Abbreviations

AD: Alzheimer disease
ADNI: Alzheimer's Disease Neuroimaging Initiative
ANN: artificial neural network
AUROC: area under the receiver operating characteristic curve
CAD: computer-aided diagnosis
FDG: fluorodeoxyglucose
FFNN: feedforward neural network
FN: false negative
FP: false positive
HC: healthy control
LDA: linear discriminant analysis
ML: machine learning
MRI: magnetic resonance imagining
N3: nonparametric nonuniform intensity normalization
OASIS: Open Access Series of Imaging Studies
PCA: principal component analysis
PET: positron emission tomography
sMRI: structural magnetic resonance imaging
SPM: statistical parametric mapping
SVM: support vector machine
t-SNE: t-distributed stochastic neighbor embedding
TN: true negative
TP: true positive
VAF: voxels-as-features
ViT: vision transformer

Edited by CN Hang; submitted 23.05.24; peer-reviewed by Anonymous, M Khani, Anonymous; revised version received 09.02.25; accepted 10.02.25; published 21.04.25.

Please cite as:
Lazli L
Improved Alzheimer Disease Diagnosis With a Machine Learning Approach and Neuroimaging: Case Study Development
JMIRx Med 2025;6:e60866
URL: https://xmed.jmir.org/2025/1/e60866
doi:10.2196/60866

Assessment of SARC-F Sensitivity for Probable Sarcopenia Among Community-Dwelling Older Adults: Cross-Sectional Questionnaire Study

David Propst[1*], MPAS, DMSc; Lauren Biscardi[1*], MS, MBA, PhD; Tim Dornemann[2*], MA, EdD

[1]Department of Exercise Science, School of Health Sciences, Barton College, 200 Acc Dr W, Wilson, NC, United States

[2]Department of Exercise Science, School of Health Sciences, North Carolina Wesleyan College, Rocky Mount, NC, United States

[*]all authors contributed equally

Corresponding Author:
David Propst, MPAS, DMSc
Department of Exercise Science, School of Health Sciences, Barton College, 200 Acc Dr W, Wilson, NC, United States

Related Articles:

Companion article: https://www.medrxiv.org/content/10.1101/2023.10.31.23297840v2

Companion article: https://med.jmirx.org/2025/1/e78552

Companion article: https://med.jmirx.org/2025/1/e77582

Companion article: https://med.jmirx.org/2025/1/e77497

Abstract

Background: The European Working Group on Sarcopenia in Older People (EWGSOP2) recommends the use of the 5-item SARC-F (strength, assistance with walking, rising from a chair, climbing stairs, and falls) questionnaire by clinicians to screen for probable sarcopenia. The recommended threshold of ≥4 has low sensitivity and high specificity in identifying probable sarcopenia. While this high threshold is effective in excluding clients without probable sarcopenia, challenges exist in using this screening tool to identify clients with low muscle strength.

Objective: This study aims to reassess the use of SARC-F in a primary care clinic for the determination of incidence of probable sarcopenia and to evaluate if a handgrip strength test is necessary for its diagnosis.

Methods: We screened 204 patients aged ≥65 years (117 men and 87 women) during routine visits with the SARC-F questionnaire. Probable sarcopenia was defined by EWGSOP2 grip strength cut points (≤27 kg for men and ≤16 kg for women). Receiver operating characteristic analysis was performed to identify the SARC-F threshold that best balanced sensitivity and specificity.

Results: Probable sarcopenia was present in 12% (n=24) of participants. The mean age (73.9, SD 6.2 years) and mean BMI (29.5, SD 5.8 kg/m²) did not differ significantly by sex; however, men showed a higher mean grip strength (36.3, SD 8.1 kg vs 22.4, SD 5.5 kg; P<.001) and lower mean SARC-F scores (0.9, SD 1.7 vs 1.9, SD 2.3; P<.001). A SARC-F cut point of ≥2 yielded an area under the curve of 0.77 (95% CI 0.67 - 0.88), with sensitivity of 0.78, specificity of 0.75, accuracy of 0.77, positive predictive value of 0.31, and negative predictive value of 0.96. The grip strength differed significantly between screen-positive and screen-negative groups at both the ≥2 and ≥4 thresholds (P<.001).

Conclusions: A SARC- F threshold of ≥2 is recommended as an optimal trade-off between sensitivity and specificity for identifying community-dwelling older adults with probable sarcopenia. This threshold is lower than the currently accepted recommendation of ≥4. Our findings promote the recommendations for early detection and treatment by medical professionals following the EWGSOP2 by improving the ability of clinicians to identify individuals with low muscle strength using this screening procedure.

(JMIRx Med 2025;6:e54475) doi:10.2196/54475

KEYWORDS

sarcopenia; neuromuscular; screening; community; scale; measure; questionnaires; diagnosis; gerontology; older adults; muscular

Introduction

Sarcopenia has been defined as a progressive loss of muscle mass and strength that adversely affects mobility, function, fall risk, and mortality in older adults [1-3]. Age-related muscle and strength loss can begin as early as 30 years of age and accelerate after 50 years of age [3-5]. The severity of muscle mass and strength loss in sarcopenia has been shown to be associated with a decreased ability to complete activities of daily living, lower quality of life, and substantially higher health care costs [5,6].

In 2018, the second European Working Group on Sarcopenia in Older People (EWGSOP2) defined a multifactorial approach to identifying sarcopenia by finding, assessing, confirming, and testing for severity [6]. This model initially screens for sarcopenia through the use of strength, assistance with walking, rising from a chair, climbing stairs, and falls through use of a clinical symptom index (eg, SARC-F [strength, assistance with walking, rising from a chair, climbing stairs, and falls]) questionnaire or using clinical suspicion [6,7].

Individuals that are identified as potentially having sarcopenia through screening undergo a muscular strength test. If strength levels meet the criteria for sarcopenia, muscle quality testing is conducted to confirm the diagnosis [3,6]. Next, the severity of sarcopenia is determined using a physical performance test [3,6].

Despite Rosenberg [8] coining the term “sarcopenia” in 1989 and the development of the *ICD-10* code *M62.84* in 2016 [9], a recent survey found that only 20% of doctors are aware of sarcopenia, a condition that can lead to falls, fractures, disability, and chronic diseases [10]. If physicians are not aware of sarcopenia, they may not screen for it or diagnose it correctly. This can lead to delays in treatment, which can have serious consequences for patients.

Early detection of sarcopenia through screening programs is crucial, as evidenced by research demonstrating that screening can lead to increased quality-adjusted life years and improved health outcomes for older adults [11,12].

While research has been conducted on various aspects of sarcopenia, including its prevalence, risk factors, and health outcomes, there has been limited focus on the practical challenges of managing this condition in primary care settings. This gap in the literature is concerning, given that primary care serves as the first point of contact for patients and plays a crucial role in early detection and management of sarcopenia [12]. Diagnosis of sarcopenia requires muscle strength testing, muscle quality testing, and a physical performance test, which is not practical in a primary care clinic. A recent review by Porter et al [13] found that primary care providers were estimated to require 26.7 hours per day, comprising 14.1 hours per day for preventive care, 7.2 hours per day for chronic disease care, 2.2 hours per day for acute care, and 3.2 hours per day for documentation and inbox management. Therefore, any additional screening must demonstrate accuracy along with being both time-efficient and cost-effective.

The EWGSOP2 pathway classifies patients as having probable sarcopenia when a brief symptom screen (eg, SARC-F) is followed by objectively low muscle strength (ie, grip or chair-stand) [6]. To embed this approach in routine care, our clinic now screens every patient aged ≥65 years during the annual physical examination. Using these real-world data, we posed two questions: (1) How common is probable sarcopenia in our practice? and (2) Can the SARC-F alone, with an optimized cut-point, serve as an efficient first-line screen? Prior studies have linked SARC-F to grip strength but did not validate a lower threshold in primary care.

Methods

A total of 204 community-dwelling older adults (ie, 87 female and 117 males) 65 years or older were screened during their regularly scheduled physician visits. Participants completed a SARC-F questionnaire and a grip strength assessment. Participant demographic data including age, gender, and BMI were recorded.

Inclusion and Exclusion Criteria

Community-dwelling adults aged ≥65 years who attended routine primary care appointments between November 2022 and March 2023 and were able to complete the SARC-F questionnaire and the 3-trial dominant-hand grip-strength test were eligible for inclusion. Patients were excluded if acute illness, recent upper-limb injury, severe arthritis, neurologic disease, or marked cognitive impairment precluded safe grip testing or questionnaire completion. These criteria reflect pragmatic screening practices and maximize both patient safety and data validity.

SARC-F Questionnaire

The SARC-F was selected as the screening tool of interest in this study. The SARC-F is a five question self-report survey developed by Malmstrom et al [7] to detect clinical symptoms of sarcopenia. The SARC-F questions include asking the patients to report difficulties with strength, assistance walking, rising from a chair, climbing stairs, and falls. The first four items are scored as 0 (no difficulty), 1 (some difficulty), or 2 (a lot of difficulty). Number of falls in the past year is rated as 0 (no falls), 1 (between 1 - 3 falls), or 2 (4 or more falls). The sensitivity is low to moderate, and the specificity is high to predict low muscle strength when a cutoff value of ≥4 is used.

Grip Strength

Muscle strength is the criterion used to detect probable sarcopenia in clinical settings [6]. Grip strength was selected as the measure of skeletal muscle strength because it is a quick and easy tool to administer during physician visits. Diagnosis of probable sarcopenia was assessed using the gender-specific recommended cutoff values for grip strength by the EWGSOP2 [6,14]. These values are <27 kg for men and <16 kg for women [14]. All grip tests were performed in private exam rooms by the first author. Participants sat with elbows flexed at 90°, wrists in a neutral position, and feet flat. Using a calibrated digital dynamometer (Sutekus Digital), each participant performed three maximal efforts (3 - 5 s) with 30 - 60 seconds of rest. The highest value for the dominant hand was used for analysis.

Ethical Considerations

This study was approved by the Barton College Institutional Review Board (IRB #2022000034; approval date January 25, 2023). As SARC-F screening and grip-strength testing are standard components of routine visits for adults 65 years or older at the study clinic, informed consent was not required as the data were obtained from deidentified medical records in accordance with the Health Insurance Portability and Accountability Act. All patient information was anonymized prior to analysis to ensure confidentiality. The collected data were anonymized, and no compensation was provided to participants.

Data Collection and Statistical Analysis

Deidentified encounter records supplied data on age, sex, BMI, SARC-F score, and dominant-hand grip strength. Normality was assessed using Kolmogorov-Smirnov tests and histograms. Between-group differences were analyzed with independent 2-tailed *t* tests (parametric) or Mann-Whitney *U* tests (nonparametric). Receiver operating characteristic (ROC) analysis evaluated the ability of SARC-F to detect probable sarcopenia (EWGSOP2 grip-strength thresholds) and generated area under the curve (AUC) estimates with 95% CIs. Sensitivity, specificity, predictive values, and accuracy were calculated at cut points 2 and 4. Effect sizes (Cohen *d* or *r*) quantified the magnitude of differences. Post hoc power for the ROC (n=204; AUC=0.75) was 98.6%.

A ROC curve was used to determine a threshold (SARC-F score) that optimized the balance between sensitivity and specificity for diagnosing probable sarcopenia. The AUC was calculated to present the ability of the SARC-F score to discriminate between probable sarcopenic and nonsarcopenic individuals. An AUC of 1.0 indicates perfect discrimination capability, 0.5 indicates discrimination capability equal to that of chance, and 0.0 indicates that all subjects are incorrectly classified.

Sensitivity, specificity, positive predictive value, negative predictive value, and false positive rate were calculated for SARC-F threshold scores. Sensitivity was calculated as the number of participants diagnosed with probable sarcopenia that were correctly identified by the SARC-F screening. Specificity was calculated as the number of participants not diagnosed with probable sarcopenia that correctly screened negative with the SARC-F. Positive predictive value was calculated as the number of participants diagnosed with probable sarcopenia that screened positive with the SARC-F. Negative predictive value was calculated as the number of participants without probable sarcopenia that screened negative with the SARC-F. The false positive rate was calculated as the ratio of the number of participants screened positive by the SARC-F without probable sarcopenia to the number of participants who were not diagnosed with probable sarcopenia. Accuracy was also calculated at each SARC-F threshold as the proportion of correctly classified patients (both true positives and true negatives).

Comparisons of muscle strength between groups determined by the SARC-F threshold of 2 and previously recommended SARC-F threshold of 4 were performed, following the between-group comparison procedures listed above. When variances were not equal, Welch *t* test was used. The α was set at .05. All statistical analyses were performed using RStudio software (version 2023.06.1; Posit PBC).

Results

SARC-F Questionnaire Scores

Probable sarcopenia was present in 12% (n=24) of participants. Participant characteristics for age, BMI, grip strength, and SARC-F score are presented in Table 1. There was a significant difference in grip strength between men and women ($t_{99.51}$=−14.25; *P*<.001; *d*=1.95) and SARC-F score (*U*=6307; *P*<.001; *r*=0.24). The sex-specific distribution of SARC-F scores is illustrated in Figure 1. There was no significant difference between BMI of men and women ($t_{145.3}$=1.39; *P*=.17) or age (t_{201}=−0.134; *P*=.89).

Table . Participant characteristics (N=204).

Variables	Overall (N=204), mean (SD)	Women (n=87), median (SD)	Men (n=117), median (SD)	*P* value
Age (years)	73.9 (6.2)	73.8 (5.9)	73.9 (6.4)	.89
BMI (kg/m^2)	29.5 (5.8)	30.2 (6.8)	29.0 (4.8)	.17
Grip strength (kg)	30.4 (9.9)	22.4 (5.5)	36.3 (8.1)	<.001[a]
SARC-F[b] score	1.33 (2.01)	1.88 (2.31)	0.92 (1.65)	<.001[a]

[a]Denotes significant difference from females (*P*<.001).

[b]SARC-F: strength, assistance with walking, rising from a chair, climbing stairs, and falls.

Figure 1. Distribution of SARC-F scores by sex. Histograms show score frequencies for male (left) and female participants (right), respectively. SARC-F: strength, assistance with walking, rising from a chair, climbing stairs, and falls.

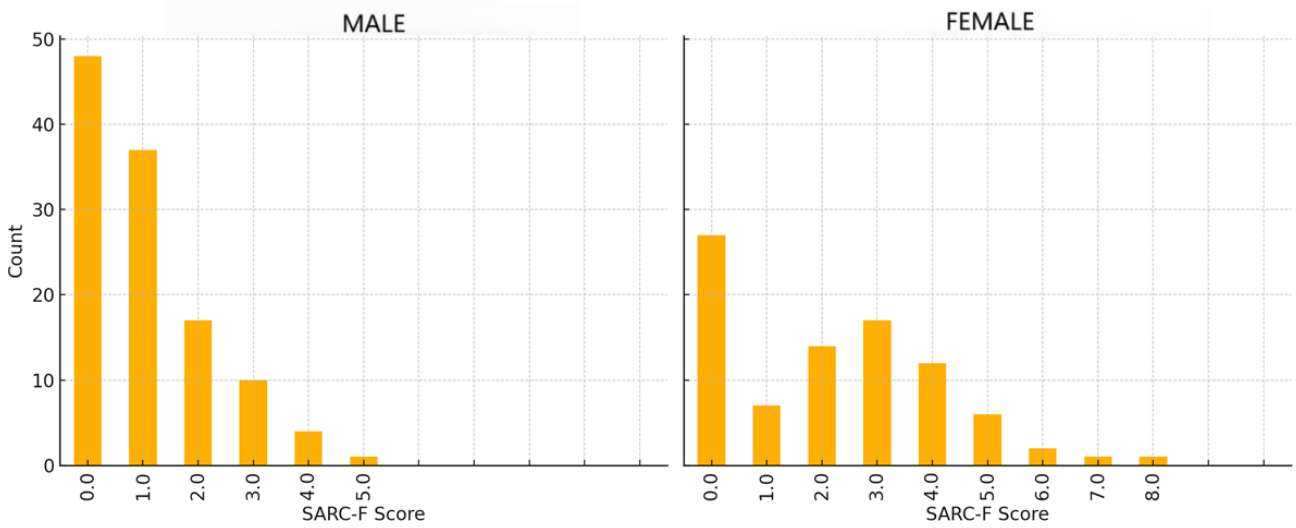


Figure 2 presents the combined ROC curve for thresholds ≥2 and ≥4. The AUC for both thresholds was 0.752 (95% CI 0.66-0.84). We compared the diagnostic performance of SARC-F across the two commonly used cut points (≥2 vs ≥4). Using DeLong test for paired ROC curves, the AUCs were not significantly different (AUC 0.752 vs 0.752; *P*=.98), supporting the clinical preference for the more sensitive ≥2 threshold. A post hoc power analysis for the ROC curve revealed statistical power of 99.5%. Calculations for sensitivity, specificity, positive predictive value, negative predictive value, false positive rate, and accuracy for SARC-F cutoff scores are presented in Table 2.

Figure 2. Combined ROC curves for SARC-F thresholds ≥2 and ≥4. AUC: area under the curve; ROC: receiver operating characteristic; SARC-F: strength, assistance with walking, rising from a chair, climbing stairs, and falls.

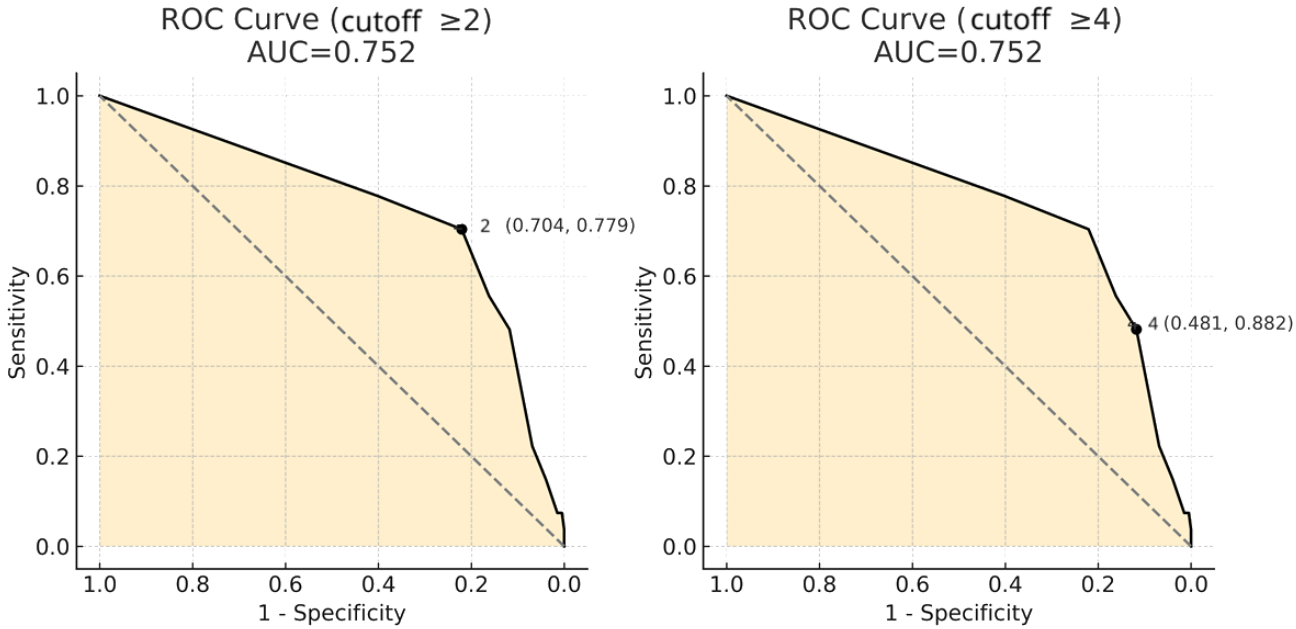


Table . Diagnostic operating characteristics at SARC-F[a] thresholds.

Cutoff values	Sensitivity (95% CI)	Specificity (95% CI)	FPR[b]	PPV[c]	NPV[d]	Accuracy
0	1.00 (1.00 - 1.00)	0.00 (0.00 - 0.00)	1.00	0.12	1.00	0.12
1	0.79 (0.63 - 0.96)	0.60 (0.53 - 0.67)	0.40	0.21	1.00	0.62
2	0.75 (0.58 - 0.92)	0.78 (0.72 - 0.83)	0.22	0.31	0.96	0.77
3	0.63 (0.42 - 0.83)	0.83 (0.78 - 0.88)	0.17	0.33	0.94	0.81
4	0.58 (0.38 - 0.75)	0.88 (0.83 - 0.93)	0.12	0.40	0.94	0.85
5	0.29 (0.13 - 0.46)	0.93 (0.89 - 0.96)	0.07	0.35	0.91	0.85
6	0.17 (0.04 - 0.33)	0.96 (0.96 - 0.98)	0.04	0.33	0.90	0.86
7	0.08 (0.00 - 0.21)	0.98 (0.96 - 1.00)	0.02	0.40	0.89	0.88
8	0.08 (0.00 - 0.21)	0.99 (0.98 - 1.00)	0.01	0.67	0.89	0.89

[a]SARC-F: strength, assistance with walking, rising from a chair, climbing stairs, and falls.

[b]FPR: false positive rate.

[c]PPV: positive predictive value.

[d]NPV: negative predictive value.

Grip Strength

Using a SARC-F cut point of ≥2, participants classified as having probable sarcopenia (n=64) had higher SARC-F scores (t_{59}=16.7; P<.001), were older (t_{80}=3.3; P=.001), had higher BMI ($t_{76.9}$= 2.7; P=.009), and demonstrated lower grip strength (t_{121}=8.0; P<.001) than those with SARC-F <2. Using a cut point of ≥4, SARC-F scores and grip strength differed significantly ($t_{66.9}$= 7.8; P<.001), whereas age and BMI were similar (P=.05 and P=.06, respectively). Mean values are summarized in Table 3.

Table . Comparison of demographic and strength variables by SARC-F threshold.

SARC-F[a] threshold	Participants, n	SARC-F, mean (SD)	Age (years), mean (SD)	BMI (kg/m²), mean (SD)	Grip strength (kg), mean (SD)
<2	167	0.23 (0.4)	72.8 (5.3)	28.9 (4.8)	33.5 (9.2)
≥2	64	3.97 (1.8)	76.7 (7.4)	31.4 (7.4)	23.5 (7.7)
<4	194	0.54 (0.9)	73.3 (5.7)	29.1 (5.3)	32.4 (9.6)
≥4	37	5.11 (1.4)	76.9 (7.9)	32.0 (7.3)	22.0 (6.4)

[a]SARC-F: strength, assistance with walking, rising from a chair, climbing stairs, and falls.

The composite ROC curve for the two cut points is presented in Figure 2. The AUC for both thresholds was 0.752 (95% CI 0.66 - 0.84). DeLong test showed no significant difference between AUCs (P=.98), supporting the clinical use of the more sensitive ≥2 threshold. Post hoc power for the ROC analysis was 99.5%. Diagnostic operating characteristics for each threshold are provided in Table 2.

Discussion

Principal Findings

A SARC-F cut point of ≥2 balanced sensitivity and specificity better than the traditional ≥4 threshold, identifying probable sarcopenia in 31% (n=63) of community-dwelling adults 65 years or older without adding clinic burden. Men demonstrated higher grip strength and lower SARC-F scores than women, reaffirming sex-specific muscle-strength disparities.

Comparison With Prior Work

Earlier studies reported high specificity but modest sensitivity when applying a SARC-F ≥4 [6,7,15]; our findings replicated this pattern (58% sensitivity, 88% specificity) while confirming that lowering the threshold to ≥2 improves case finding (78% sensitivity) while maintaining acceptable specificity (75%). Our AUC of 0.75 aligns with the AUC of 0.71 as reported by Erbas Sacar et al [16], supporting the tool's value as a screening and not a stand-alone diagnostic test. Recent authors have advocated thresholds as low as ≥1 for maximal sensitivity [14,16,17]; our operating characteristic table (Table 2) illustrates the same trade-off: as the cut point increases, sensitivity decreases and specificity increases. DeLong test showed no difference between AUCs for the two thresholds (P=.98), strengthening the argument for the more sensitive ≥2 cut point in primary care.

Strengths and Limitations

First, real-world implementation during annual visits increases external validity. Second, standardized grip-strength testing

minimized measurement error. Third, the sample (N=204) provided 98.6% post hoc power for ROC analyses. Limitations included (1) the single-site, cross-sectional design limits generalizability and causal inference; (2) SARC-F relies on self-report and may incur recall bias; (3) potential confounders (physical activity, cognitive status, comorbidities) were not captured; and (4) grip strength was measured once, and functional measures such as gait speed were unavailable. Each factor may attenuate or inflate the observed associations, underscoring the need for multimodal assessment in future research.

Future Directions

Prospective, multicenter studies should validate the ≥2 threshold across diverse settings, incorporate additional functional tests, and examine longitudinal outcomes (ie, falls, hospitalization, disability). Cost-effectiveness analyses could further justify routine SARC-F screening in primary care, and digital integration of the questionnaire into electronic health records may streamline population-level implementation.

Conclusion

A SARC-F cut point of ≥2 offers a feasible, time-efficient approach to flag older primary care patients who require confirmatory strength testing, aligning with EWGSOP2 recommendations for early clinical intervention. The tool should be used to complement—rather than replace—comprehensive diagnostic workups.

Acknowledgments

All authors declared that they had insufficient or no funding to support open access publication of this manuscript, including from affiliated organizations or institutions, funding agencies, or other organizations. JMIR Publications provided article processing fee support for the publication of this article.

Data Availability

The datasets used and analyzed during this study are not publicly available due to participant confidentiality and privacy restrictions but are available from the corresponding author upon reasonable request.

Authors' Contributions

Conceptualization: DP, TD
Data curation: DP, LB
Formal analysis: LB
Methodology: DP, LB, TD
Project administration: TD
Supervision: TD
Visualization: LB
Writing – original draft : DP, LB
Writing – review & editing: DP, LB, TD

Conflicts of Interest

None declared.

References

1. Beaudart C, Biver E, Reginster JY, et al. Validation of the SarQoL®, a specific health-related quality of life questionnaire for sarcopenia. J Cachexia Sarcopenia Muscle 2017 Apr;8(2):238-244. [doi: 10.1002/jcsm.12149] [Medline: 27897430]
2. Papadopoulou SK. Sarcopenia: a contemporary health problem among older adult populations. Nutrients 2020 May 1;12(5):1293. [doi: 10.3390/nu12051293] [Medline: 32370051]
3. Rubbieri G, Mossello E, Di Bari M. Techniques for the diagnosis of sarcopenia. Clin Cases Miner Bone Metab 2014 Sep;11(3):181-184. [Medline: 25568650]
4. Hunt D, Chapa D, Hess B, Swanick K, Hovanec A. The importance of resistance training in the treatment of sarcopenia. JNEP 2014 Sep;5(3):39-44. [doi: 10.5430/jnep.v5n3p39]
5. Jackson KL, Hunt D, Chapa D, Gropper SS. Sarcopenia-A baby boomers dilemma for nurse practitioners to discover, diagnose, and treat. JNEP 2018 Sep;8(9):77. [doi: 10.5430/jnep.v8n9p77]
6. Cruz-Jentoft AJ, Bahat G, Bauer J, et al. Sarcopenia: revised European consensus on definition and diagnosis. Age Ageing 2019 Jul 1;48(4):601-601. [doi: 10.1093/ageing/afz046]
7. Malmstrom TK, Miller DK, Simonsick EM, Ferrucci L, Morley JE. SARC-F: a symptom score to predict persons with sarcopenia at risk for poor functional outcomes. J Cachexia Sarcopenia Muscle 2016 Mar;7(1):28-36. [doi: 10.1002/jcsm.12048] [Medline: 27066316]
8. Rosenberg IH. Summary comments. Am J Clin Nutr 1989 Nov;50(5):1231-1233. [doi: 10.1093/ajcn/50.5.1231]

9. Anker SD, Morley JE, von Haehling S. Welcome to the ICD-10 code for sarcopenia. J Cachexia Sarcopenia Muscle 2016 Dec;7(5):512-514. [doi: 10.1002/jcsm.12147] [Medline: 27891296]
10. Guralnik JM, Cawthon PM, Bhasin S, et al. Limited physician knowledge of sarcopenia: a survey. J American Geriatrics Society 2023 May;71(5):1595-1602. [doi: 10.1111/jgs.18227]
11. Darvishi A, Hemami MR, Shafiee G, et al. Sarcopenia screening strategies in older people: a cost effectiveness analysis in Iran. BMC Public Health 2021 May 17;21(1):926. [doi: 10.1186/s12889-021-10511-7] [Medline: 34001057]
12. Kandayah T, Safian N, Azhar Shah S, Abdul Manaf MR. Challenges in the management of sarcopenia in the primary care setting: a scoping review. Int J Environ Res Public Health 2023 Mar 15;20(6):5179. [doi: 10.3390/ijerph20065179] [Medline: 36982085]
13. Porter J, Boyd C, Skandari MR, Laiteerapong N. Revisiting the time needed to provide adult primary care. J Gen Intern Med 2023 Jan;38(1):147-155. [doi: 10.1007/s11606-022-07707-x] [Medline: 35776372]
14. Dodds RM, Murray JC, Robinson SM, Sayer AA. The identification of probable sarcopenia in early old age based on the SARC-F tool and clinical suspicion: findings from the 1946 British birth cohort. Eur Geriatr Med 2020 Jun;11(3):433-441. [doi: 10.1007/s41999-020-00310-5] [Medline: 32297269]
15. Bahat G, Yilmaz O, Kılıç C, Oren MM, Karan MA. Performance of SARC-F in regard to sarcopenia definitions, muscle mass and functional measures. J Nutr Health Aging 2018;22(8):898-903. [doi: 10.1007/s12603-018-1067-8] [Medline: 30272090]
16. Erbas Sacar D, Kilic C, Karan MA, Bahat G. Ability of SARC-F to find probable sarcopenia cases in older adults. J Nutr Health Aging 2021;25(6):757-761. [doi: 10.1007/s12603-021-1617-3] [Medline: 34179930]
17. Du W, Gao C, Wang X, et al. Validity of the SARC-F questionnaire in assessing sarcopenia in patients with chronic kidney disease: a cross-sectional study. Front Med (Lausanne) 2023;10:1188971. [doi: 10.3389/fmed.2023.1188971] [Medline: 37534318]

Abbreviations

AUC: area under the curve
EWGSOP2: The European Working Group on Sarcopenia in Older People
ROC: receiver operating characteristic
SARC-F: strength, assistance with walking, rising from a chair, climbing stairs, and falls

Edited by A Schwartz; submitted 10.11.23; peer-reviewed by Anonymous, Anonymous; revised version received 02.05.25; accepted 12.05.25; published 25.07.25.

Please cite as:
Propst D, Biscardi L, Dornemann T
Assessment of SARC-F Sensitivity for Probable Sarcopenia Among Community-Dwelling Older Adults: Cross-Sectional Questionnaire Study
JMIRx Med 2025;6:e54475
URL: https://xmed.jmir.org/2025/1/e54475
doi:10.2196/54475

Checklist Approach to Developing and Implementing AI in Clinical Settings: Instrument Development Study

Ayomide Owoyemi[1], MScPH, MD, PhD; Joanne Osuchukwu[2], MD; Megan E Salwei[3], BSc, MSc, PhD; Andrew Boyd[1], BSc, MD

[1]Department of Biomedical and Health Informatics, University of Illinois Chicago, 1919 W Taylor, Chicago, IL, United States

[2]College of Medicine, University of Cincinnati, Cincinnati, OH, United States

[3]Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, TN, United States

Corresponding Author:
Ayomide Owoyemi, MScPH, MD, PhD
Department of Biomedical and Health Informatics, University of Illinois Chicago, 1919 W Taylor, Chicago, IL, United States

Related Articles:

Companion article: https://www.medrxiv.org/content/10.1101/2024.08.08.24311701v1

Companion article: https://med.jmirx.org/2025/1/e69869

Companion article: https://med.jmirx.org/2025/1/e70058

Companion article: https://med.jmirx.org/2025/1/e69593

Companion article: https://med.jmirx.org/2025/1/e69594

Companion article: https://med.jmirx.org/2025/1/e69870

Companion article: https://med.jmirx.org/2025/1/e69595

Companion article: https://med.jmirx.org/2025/1/e69537

Abstract

Background: The integration of artificial intelligence (AI) in health care settings demands a nuanced approach that considers both technical performance and sociotechnical factors.

Objective: This study aimed to develop a checklist that addresses the sociotechnical aspects of AI deployment in health care and provides a structured, holistic guide for teams involved in the life cycle of AI systems.

Methods: A literature synthesis identified 20 relevant studies, forming the foundation for the Clinical AI Sociotechnical Framework checklist. A modified Delphi study was then conducted with 35 global health care professionals. Participants assessed the checklist's relevance across 4 stages: "Planning," "Design," "Development," and "Proposed Implementation." A consensus threshold of 80% was established for each item. IQRs and Cronbach α were calculated to assess agreement and reliability.

Results: The initial checklist had 45 questions. Following participant feedback, the checklist was refined to 34 items, and a final round saw 100% consensus on all items (mean score >0.8, IQR 0). Based on the outcome of the Delphi study, a final checklist was outlined, with 1 more question added to make 35 questions in total.

Conclusions: The Clinical AI Sociotechnical Framework checklist provides a comprehensive, structured approach to developing and implementing AI in clinical settings, addressing technical and social factors critical for adoption and success. This checklist is a practical tool that aligns AI development with real-world clinical needs, aiming to enhance patient outcomes and integrate smoothly into health care workflows.

(JMIRx Med 2025;6:e65565) doi:10.2196/65565

KEYWORDS

artificial intelligence; machine learning; algorithm; model; analytics; AI deployment; human-AI interaction; AI integration; checklist; clinical workflow; clinical setting; literature review

Introduction

The implementation of any technology in a real-world setting, especially a clinical one, requires adequate consideration of the social aspects of its application alongside the technical considerations [1]. The National Academy of Medicine report highlighted the need to "understand the technical, cognitive, social, and political factors in play and incentives impacting integration of Artificial Intelligence (AI) into health care workflows" [2]. It is important to understand the context in which the technology will be used, how it will work with existing workflows without disruption, and how it will be accepted by the people who will have to use it. Historically, in the development of AI systems, the technical perspective has taken preeminence over how they fit and work in the real world, and this has resulted in AI systems falling short of their translational goals [3]. In general, AI tools have shown promise in development, but few have been able to translate into the real-world settings for patient management [4]. For example, for a management decision tool built and deployed in a hospital in Utah for diabetes management, there was a challenge of not offering all the information that was desired by clinicians and patients to decide on type 2 diabetes management [5].

Despite the numerous proof-of-concept publications in this field, the lack of robust frameworks for supporting the development and management of these tools has been one of the main barriers to their adoption in health care [6]. There is a paucity of specific guidance and rigorous best practices for people designing and developing AI solutions targeted at clinical settings and use cases. A review conducted by Gama et al [7] highlighted the need to develop an AI-specific implementation framework because there is an unrealized opportunity to draw insights from implementation science, as well as to use theoretical and practical insights, to accelerate and improve on the implementation of AI in clinical settings.

There have been a few frameworks and guidelines proposed recently. Salwei and Carayon [1] developed a sociotechnical systems framework for AI that acknowledges the social and technical aspects of work that relate to the successful design and implementation of AI. Their model demonstrates that an AI can only integrate into clinical workflows if it fits within the context, or the work system, in which it is implemented. The CONSORT (Consolidated Standards of Reporting Trials)-AI extension and TRIPOD (Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis) are examples of models that are narrow in their application and are focused on trials, performance, and comparison, which are only helpful in a single phase of the AI life cycle [8,9]. However, most of the existing frameworks gloss over relevant sociotechnical factors, while others only target specific stages in the AI development cycle, and almost all have no easy-to-use checklist. This study sought to develop a framework and operationalize it as a checklist that covers all the aspects of the development cycle and holistically addresses sociotechnical factors across those phases.

Methods

Literature Synthesis

We conducted a literature search on the MEDLINE via OVID and Embase databases between June 25 and 30, 2023. Our search focused on studies examining AI in clinical settings, particularly those addressing frameworks, guidelines, and theories for AI implementation, design, and evaluation. The following keywords were used in the search: "Artificial intelligence," "Framework," "Guideline," "Theory," "Implementation," "Evaluation," "Design," "Development," "Clinical Settings," "Clinical Care," "Hospital," "Clinic," and "Patient Care." There were no restrictions on the publication dates of the studies, meaning articles from any year were considered in the search. This initial search identified 573 potential studies. We screened the abstracts of these studies using the following inclusion criteria:

- Studies involving the application of AI by health care providers in a clinical setting
- Research that used a conceptual or theoretical framework related to AI in clinical care
- Primary qualitative studies that focused on the design, implementation, or evaluation of AI in clinical care, regardless of whether a distinct framework was used

We excluded studies that:

- Focused primarily on patient-related outcomes
- Concentrated on the technical or computational aspects of AI without clinical integration

We identified 19 relevant studies for full-text review. Three were excluded (one reporting guideline, one study protocol, and one commentary). Through citation tracking, we added 4 additional relevant studies, bringing the final sample to 20 articles. These 20 studies were thoroughly reviewed, and key points, themes, and insights were extracted. We then synthesized these insights with findings from a previously conducted primary study [10] on the implementation and user experience of an AI-powered sepsis alert system. Using a mind map approach, we organized the themes and insights into key domains to develop our framework.

The Modified Delphi Study

The framework developed from the literature synthesis was used to develop a preliminary draft of a checklist targeted at supporting teams designing and developing AI systems for clinical settings. This draft was shared with selected experts for review, edits, and improvements using a Delphi method. The Delphi method is a procedure for reaching a consensus with a group of people who are typically experts on the subject through controlled assessments [11]. The technique has been used in health care to achieve consensus in establishing guidelines or

treatment protocols when evidence is limited, inadequate, or contradictory [12]. For this study, a modified approach was used, which involved the development of the initial checklist questions by the researcher rather than the panelists. This approach ensured that the questions were grounded in the literature framework and leveraged the researcher's expertise. This modification helped streamline the process and ensure that the questions were relevant to the specific context of AI system development in clinical settings. The panelists were then asked to refine and validate these questions, rather than generating them from scratch.

The modified Delphi study was conducted between January 23 and March 14, 2024. The selection of Delphi panelists followed a process aimed at ensuring diversity in expertise and professional background. Potential participants were recruited through targeted outreach on platforms such as email listserves, LinkedIn, Twitter, and closed WhatsApp groups. To be eligible, participants were required to hold advanced degrees and have at least 2 years of professional experience in fields directly related to AI systems in health care. Specifically, panelists were selected based on their expertise in areas such as medicine (doctors and nurses), health informatics, AI research, AI engineering, health care administration, human factors research, health care system research, implementation science, health care product management, health ethics, and safety. The global nature of the study welcomed participants from any country, ensuring a broad range of perspectives.

Interested individuals were initially asked to complete a preliminary form to provide background information about their experience and qualifications. This form was used to filter suitable candidates for inclusion in the Delphi panel. Invitations were then sent to selected candidates, along with a detailed information letter explaining the study's goals and procedures. A pretest was conducted with a panel comprising 5 professionals, each with some expertise in the fields of health care and technology. Their feedback helped refine the checklist to ensure clarity, making it easier for participants to understand and respond accurately.

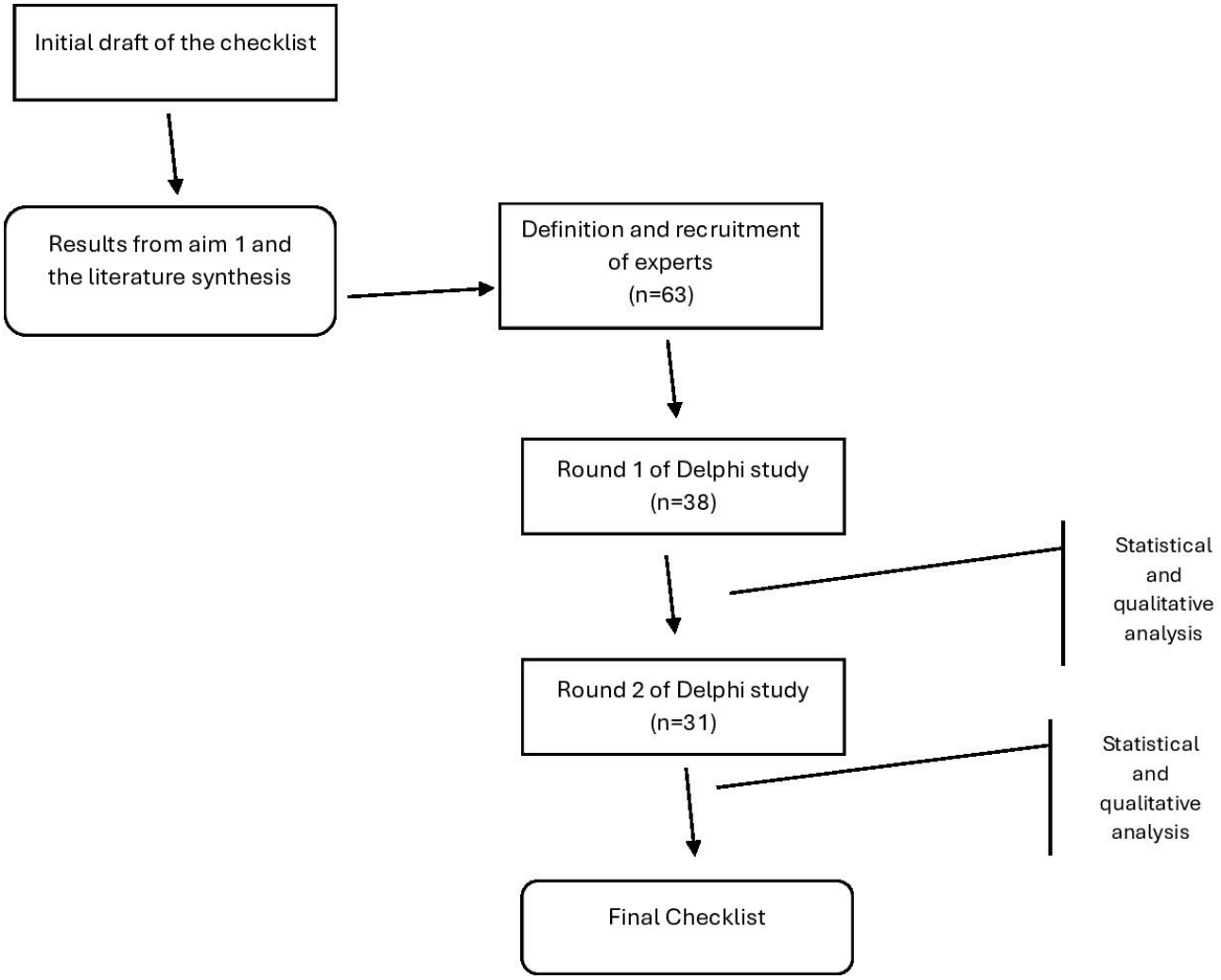
Participants who agreed to take part accessed the first round of the Delphi survey through a link in the email, which led to the consent form and survey. Data collection was done using Google Forms. To avoid bias, the panelists remained anonymous to each other throughout the process.

The preliminary survey comprised 45 questions designed to assess the relevance of each checklist item to the AI system's design and development process. A Likert scale from 1 ("Not Relevant") to 5 ("Highly Relevant") was used, along with open-ended comment fields for feedback and suggestions. The checklist was organized into four stages of AI system development: (1) planning, (2) design, (3) development, and (4) proposed implementation. Each stage aligned with 1 of the 6 domains in our framework.

After completion of the preliminary survey, the results were analyzed to assess the level of consensus among panelists. Based on the analysis, along with participants' feedback and comments, the checklist was revised and updated for the second round of the Delphi process. All the initial panelists were also invited for the second round even if they missed the first. This approach was based on the study by Boel et al [13], which showed that inviting panel members who missed a previous round to a subsequent round led to better representations of opinions and reduced the chances of false consensus while not influencing the outcome. The results of the analysis and feedback were added to the questionnaire for the second round. The whole process is highlighted in Figure 1.

Questions rated 4 or higher were classified as "relevant" to streamline the analysis. At the same time, those rated 3 or lower were deemed "irrelevant." This categorization facilitated a more efficient evaluation of the panelists' responses. Descriptive statistics were used to analyze the results of each round, along with an analysis of the IQR for each question. In determining the threshold for consensus among panelists, a mean score of 0.8 (representing 80% agreement) was established a priori as the benchmark. Questions with a mean score above 0.8 and an IQR of 0 were deemed to have consensus among the participants. Lastly, the Cronbach α reliability coefficient was calculated to evaluate the interitem reliability. The qualitative data collected during each round were analyzed using inductive content analysis. Quantitative analyses were conducted using the Python programming language in JupyterLab for Windows (Project Jupyter).

Figure 1. The process of developing the checklist.



Ethical Considerations

This study was conducted in accordance with institutional ethical guidelines for research involving human subjects and was approved by the University of Illinois Chicago Institutional Review Board under protocol STUDY2023-0535-MOD003. Participants provided informed consent, ensuring they were aware of the study's purpose, procedures, potential risks, and their right to withdraw at any time. All data collected were either anonymized or deidentified to protect participant privacy, with strict safeguards in place to ensure confidentiality. Additionally, no financial or material compensation was provided to participants in this Delphi study, and participation was entirely voluntary.

Results

Literature Synthesis

The literature search identified 20 studies [1,3,7,14-30] that proposed a framework, guideline, or approach for the design, development, implementation, or evaluation of AI for clinical use cases (Figure 2). A total of 14 (65%) of these addressed specific areas in the AI development cycle, from design to maintenance and management, while some cut across every aspect of the cycle. The results of the literature search were synthesized with the primary research and connected using a mind map to arrive at the domains of the Clinical AI Sociotechnical Framework (CASoF), which is a sociotechnical framework to support the planning, design, development, and proposed implementation of AI systems to help better plan and predict the likely success of the AI system (Figure 3).

Figure 2. PRISMA (Preferred Reporting Items for Systematic Reviews for Meta-Analyses) flowchart.

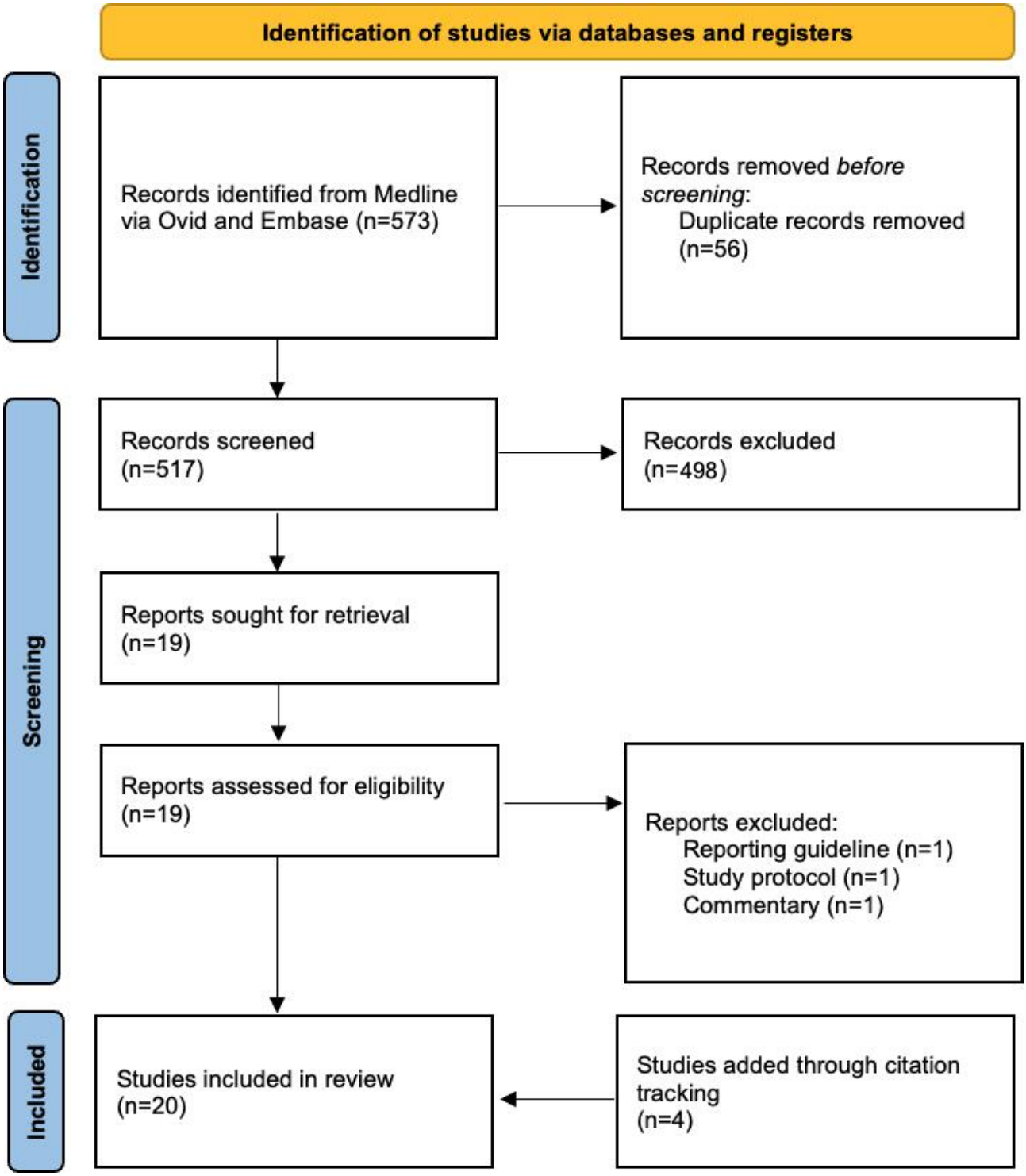
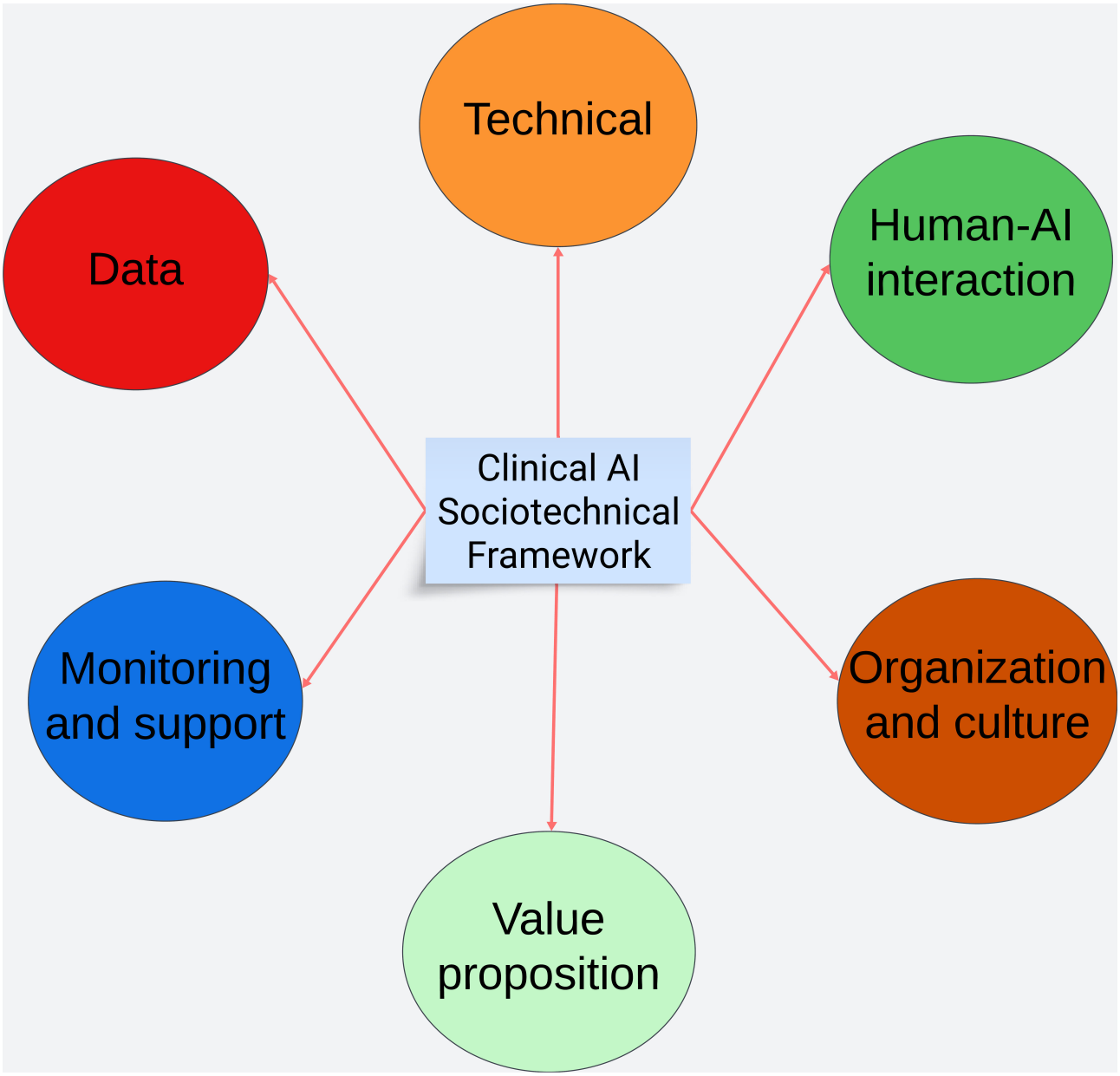


Figure 3. The Clinical AI Sociotechnical Framework. AI: artificial intelligence.



The Modified Delphi Study

Based on the CASoF, the first draft of the checklist was developed, which was shared with a team of panelists for evaluation and review using a Delphi approach. A total of 65 panelists were recruited: 21 (32%) doctors, 10 (15%) health care experts or researchers, 9 (12%) AI researchers, 4 (6%) health informaticians, 4 (6%) nurses, and 18 (28%) other professionals. Of the 65 panelists invited to participate in the study, 35 (54%) of them completed the first round of Delphi. The initial checklist had 4 overall categories that corresponded to the 4 stages in the development and deployment process, with 15 subcategories that corresponded to the domains of the CASoF that were important in each of the stages. The stages were "Planning," "Design," "Development," and "Proposed Implementation." As part of the questionnaire, panelists were asked 2 open-ended questions at the end of each of the subcategories: "Would you reframe any of the questions above?" and "Are there questions that you would add or remove from this segment?" During the first round of the Delphi, panelists suggested multiple edits and additions to the checklist. This suggested editing included the need to reframe some of the questions to make them more appropriate and clearer for a checklist. In one of the subcategories, one panelist responded as follows:

> *The last question says, "data processing." That comes across as ambiguous. What does that refer to? who will be the audience for this survey? will they understand what that means? Are we trying to abstract curation, cleaning etc into abstraction?*

At the end of the survey, panelists were asked why they might not use the checklist, and some of the responses included the following:

> *I think the checklist is long. The challenge when you have checklists this long is that people tend to gloss over them and are not intentional about answering the questions in a detailed way.*

> *Might be helpful to shorted and make more actionable. eg, policies and procedures document has been completed versus have you considered a place for policies.*
>
> *The checklist is somewhat burdensome on the AI vendor and health system. I would cut the questions in half.*

These open-ended questions were analyzed using a content analysis approach to bring out the recurrent themes and perspectives shared by the panelists in reforming and improving the questionnaire. Quantitative analyses were done, which showed a high level of agreement and relevance across most questions. Descriptive analysis was done: the mean score for the relevance of the questions on the survey exceeded 0.8 on all but one, indicating that at least 80% of respondents found the questions pertinent to their work and the topic at hand. Furthermore, the IQR was calculated to be 0 for all questions except 3, highlighting a level of consensus among respondents. The consensus and the structure of the checklist are shown in Multimedia Appendix 1.

Based on the results, comments, and feedback from the panelists, the checklist was revised. The "Design" and "Development" stages were merged into a single stage, and the "People" and "Organization and Culture" domains were merged into a single domain. The "User Experience and Workflow" and "Clinical Utility" domains were merged to create a new domain called "Human-AI Interaction." The total number of questions was reduced from 45 questions to 34 questions to make it less cumbersome and more focused. These 34 questions were sent to all the registered panelists for a second round of the Delphi process. All the recruited panelists were included in the second round and invited to review the updated checklist. Quantitative analyses were done, which showed a high level of agreement and relevance across most questions. Descriptive analysis was done: the mean score for the relevance of the questions was more than 0.8 on all questions, indicating that at least 80% of respondents found the questions pertinent to their work and the topic at hand. Furthermore, the IQR was calculated to be 0 for all questions, highlighting a level of consensus among respondents. Based on the outcome of the Delphi study, a final checklist was outlined, with 1 more question added to make 35 questions in total (Table 1).

Table . Final draft of the Clinical AI[a] Sociotechnical Framework (CASoF) checklist.

Stage and domain		Questions
Planning		
	Value proposition and utility	• Have you outlined the expected impacts on patient outcomes? • Have you outlined its expected impact on care provider efficiency and outcomes? • Has any economic analysis been conducted for the AI system?
	Data	• Have you engaged in the use of any ethical data checklist during your data collection and preparation? • Have you engaged domain experts in the data preparation, cleaning, and engineering process? • Have you delineated an approach to maintain data quality, integrity, and security?
	People, organization, and culture	• Have you identified key stakeholders and their needs? • Have you identified potential resistance or barriers within the organization? • Are there strategies in place to facilitate and ensure end-user engagement in the design and development phase? • Do you have a good understanding of the culture within the institution and changes that might be needed?
Design and development		
	Technical	• Are you planning for hardware/software (EHR[b]) systems and requirements? • Have you conducted a real-world evaluation of the model? • Are you creating support documentation for users and management, eg, model details, explainability details, data details, metrics, manuals, etc? • Have you validated clinical accuracy and reliability? • Have you secured any required regulatory approval? • Have you taken active steps to mitigate against biased results?
	Human-AI integration	• Have you conducted a simulation with end users in real work system scenarios? • Have you evaluated if the outputs are clear and understandable for the users? • Have you implemented any patient and user safety measures? • Have you accounted for and evaluated existing clinical workflows? • Are you aligning the solution with existing protocols? • Have you assessed the impact on the delivery of clinical tasks? • Have you involved and tested with users? • Has any resistance to the use of the AI system been identified and addressed? • Are you developing strategies to ensure that the alerts from the AI system are relevant, timely, and not overwhelming, to avoid alert fatigue?

Stage and domain		Questions
	Data	• Have you tested your method on various types of data to make sure it works well in different situations? • Have you planned for data drift and shift (changes in the data over time)?
Proposed implementation		
	People, organization, and culture	• Have you ensured that this intervention aligns with the existing governance and regulatory frameworks of the organization? • Have you prepared necessary training/resources for end users? • Have you considered steps to help address end users' questions and alleviate their concerns?
	Technical	• Are you planning for pilot/silent tests? • Are you providing user tools for continuous validation and evaluation of the system?
	Monitoring and support	• Have you created a plan to evaluate the success of the implementation? • Have you planned for continuous user feedback on the system? • Have you planned for regular audits, reviews, and updates? • Have you planned for continuous education and support for users?

[a]AI: artificial intelligence.

[b]EHR: electronic health record.

Discussion

Principal Findings

We introduce the CASoF checklist, which is a checklist that was developed from the results of primary studies, a literature synthesis, and a modified Delphi process that involved multiple experts and health care professionals. The CASoF, based on its sociotechnical perspective, encompasses different existing frameworks by providing a structured overview of the critical issues related to the integration, validation, and operationalization of AI in health care. The CASoF offers a high-level approach to solving the translation and adoption problems bedeviling AI systems designed for clinical settings. The CASoF can be used singly or in combination with some of the other existing frameworks in evaluating AI systems. The Diagnostic Quality Model by Lennerz et al [16] and the Clinical Explainable AI Guidelines by Jin et al [17] address diagnostic quality and explainability within medical imaging. They provide structured methodologies that could refine the CASoF by integrating rigorous quality assessments and enhancing transparency in AI tools. The strengths of these frameworks lie in their focused criteria, which could synergistically enrich the CASoF's scope, ensuring that AI's clinical implementation is both effective and sociotechnically sound.

At the end of the Delphi study and reviews, 35 final questions were agreed on based on the consensus from the panel members. Adjustments and rearrangements were made to the sequence of questions based on the comments made as part of the feedback during the Delphi study. This is the first checklist that addresses sociotechnical factors across the phases of the AI cycle with a general approach that is not limited to any specific condition or use case in clinical care. The checklist aims to help ensure that AI solutions for clinical use cases are better built for impact, adoption, and success.

The checklist focuses on sociotechnical factors most relevant to achieving these outcomes. Some of the comments by the respondents highlighted how the high-level design of the checklist was a reason they might not use it; however, the checklist is intentionally made high level to make it as brief and less cumbersome as possible. One of the reasons it is high level is to make it easy to apply quickly by designers, developers, AI engineers, informaticians clinicians, and health care organization managers for the needed assessments; therefore, this checklist should be considered as a form of minimum guideline in the development and implementation of AI systems meant for clinical settings.

The checklist is divided into 3 stages corresponding to the phases of the AI development cycle. The domains are drawn from the domains of the CASoF, which are "Value Proposition," "Data," "Human-AI Interaction," "Organization and Culture," "Technical," and "Monitoring and Support" [31]. These domains are allocated to each stage based on their relevance to that stage. Some domains recur in different stages, like "Data," "Human-AI Interaction," "Organization and Culture," and "Technical." Other domains like "Value Proposition" and "Monitoring and Support" only appear in a single phase. Questions are outlined

under each domain based on the stage they belong to. The number of questions varies per stage and domain.

The questions must be answered with a "Yes," "No," or "Partially Done." Each stage is meant to be done before and after each corresponding phase of the development cycle, so that the development team knows what to plan for and later review what has been accomplished. The "Planning" stage addresses the decision and preparation phase of the project, which is where the groundwork is laid for the subsequent design of the system. This phase involves a value proposition assessment to determine if it ensures alignment with patients' and end users' benefits. It serves to help answer a "go or no go" question across the ethical, economic, and sociotechnical dimensions of the AI tool, which is part of what the "Planning" phase in the CASoF checklist is designed to support. While the Biological-Psychological, Economic, and Social checklist by Khan and Seto [32] covers the planning aspect of AI development, it does not go beyond that phase, which is a limitation in its application.

The "Design and Development" phase covers the necessary steps and factors to be considered while building the AI system, unlike the R-AI-DIOLOGY checklist, which, apart from being focused explicitly on AI systems in radiology, only addresses the technical aspects of the design and development phases [33]. The last part of the checklist helps to plan for implementation, focusing on organization, culture, and needed monitoring. The Translational Evaluation of Healthcare AI framework checklist offers an alternative to the CASoF checklist for implementation; however, its lack of sociotechnical components, such as human-AI integration, culture and organization, and monitoring and support, which are essential for adoption and maximizing utility, is a drawback [3]. The checklist's design, development, and preimplementation aspects can also be used by payers, buyers, and decision makers to evaluate AI systems being sold or proposed to them to ensure they have been well designed and built.

Most of the existing checklists in this domain are targeted at reporting medical research carried out in AI or machine learning [34]. The CASoF checklist differs from these and other existing checklists like the Technology, Organization, and People framework–based checklist, which is focused on helping digital leaders manage adoption challenges [35]. It has no domain that addresses how the AI is designed or built, unlike the CASoF checklist. The same goes for the DECIDE-AI (Developmental and Exploratory Clinical Investigations of Decision Support Systems Driven by Artificial Intelligence) checklist, which is focused on reporting studies that involve the evaluation of AI systems during their implementation phase in the clinical setting [36]. While the CASoF checklist does not explicitly have questions that address ethical issues, there are multiple questions across different phases that raise the need to address the ethics of the data, patient outcomes, and the impact of the outputs of the AI system.

Enhancing the real-world impact of AI tools involves navigating a nuanced blend of technical and social elements. This process demands a strategic framework that guides the planning and preparation efforts throughout the AI tool's life cycle, from its initial conceptualization to its sustained application. The CASoF checklist is designed to support designers, developers, AI engineers, informaticians, clinicians, health care organization managers, and others in planning, monitoring, and evaluating AI systems being developed or sold to them for clinical care.

Limitations

While the primary research, literature synthesis, and Delphi technique offer a robust approach to the development of the framework and checklist for the development and integration of AI in the clinical setting, the real-world application could be more difficult and not as straightforward as the research might suggest. Therefore, there might be a need for continuous refinement of the CASoF through iterative feedback and broader engagement with more stakeholders. Future research should aim to include an even wider array of perspectives, particularly from underrepresented regions and specialties, to enhance the framework's comprehensiveness and applicability. The framework further encounters limitations in capturing the full spectrum of technical challenges, needs, and their implications across diverse health care contexts globally. Considering these constraints, the application of the framework will benefit from synergistic application with other existing frameworks.

Conclusion

The CASoF checklist offers an approach to bridge the gap between the technical aspects of AI and how they can be best planned to fit and work in the clinical setting, with a view to improving the impact it makes on clinical work and patient outcomes. It offers a structured strategy to mitigate challenges and obstacles in the development and implementation process. The CASoF offers an advancement over previous frameworks and approaches by holistically encapsulating the sociotechnical dimensions necessary for AI to thrive within the clinical space.

Acknowledgments

This work was supported in part by the Agency for Healthcare Research and Quality grant K01HS029042 (MES). This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Data Availability

This checklist is available in an electronic format [37].

Authors' Contributions

AO contributed to the conceptualization, data collection, formal analysis, investigation, and methodology of the study. Additionally, AO drafted the original manuscript and participated in the review and editing process. JO contributed to writing, reviewing, and editing of the manuscript. MES provided formal analysis, project administration, and supervision and contributed to the review and editing of the manuscript. AB contributed to the conceptualization and formal analysis of the study, managed the project, provided resources, and supervised the research. He also participated in the review and editing of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Summary of subcategories by domain for the first round of the Delphi study.
[DOCX File, 15 KB - xmed_v6i1e65565_app1.docx]

Checklist 1
PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews) checklist.
[DOCX File, 86 KB - xmed_v6i1e65565_app2.docx]

References

1. Salwei ME, Carayon P. A sociotechnical systems framework for the application of artificial intelligence in health care delivery. J Cogn Eng Decis Mak 2022 Dec;16(4):194-206. [doi: 10.1177/15553434221097357] [Medline: 36704421]
2. Matheny ME, Whicher D, Thadaney Israni S. Artificial intelligence in health care: a report from the National Academy of Medicine. JAMA 2020 Feb 11;323(6):509-510. [doi: 10.1001/jama.2019.21579] [Medline: 31845963]
3. Reddy S, Rogers W, Makinen VP, et al. Evaluation framework to guide implementation of AI systems into healthcare settings. BMJ Health Care Inform 2021 Oct;28(1):e100444. [doi: 10.1136/bmjhci-2021-100444] [Medline: 34642177]
4. He J, Baxter SL, Xu J, Xu J, Zhou X, Zhang K. The practical implementation of artificial intelligence technologies in medicine. Nat Med 2019 Jan;25(1):30-36. [doi: 10.1038/s41591-018-0307-0] [Medline: 30617336]
5. Tarumi S, Takeuchi W, Chalkidis G, et al. Leveraging artificial intelligence to improve chronic disease care: methods and application to pharmacotherapy decision support for type-2 diabetes mellitus. Methods Inf Med 2021 Jun;60(S 01):e32-e43. [doi: 10.1055/s-0041-1728757] [Medline: 33975376]
6. Ben-Israel D, Jacobs WB, Casha S, et al. The impact of machine learning on patient care: a systematic review. Artif Intell Med 2020 Mar;103:101785. [doi: 10.1016/j.artmed.2019.101785] [Medline: 32143792]
7. Gama F, Tyskbo D, Nygren J, Barlow J, Reed J, Svedberg P. Implementation frameworks for artificial intelligence translation into health care practice: scoping review. J Med Internet Res 2022 Jan 27;24(1):e32215. [doi: 10.2196/32215] [Medline: 35084349]
8. Liu X, Cruz Rivera S, Moher D, Calvert MJ, Denniston AK, SPIRIT-AI and CONSORT-AI Working Group. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. Nat Med 2020 Sep;26(9):1364-1374. [doi: 10.1038/s41591-020-1034-x] [Medline: 32908283]
9. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. BMJ 2015 Jan 7;350:g7594. [doi: 10.1136/bmj.g7594] [Medline: 25569120]
10. Owoyemi A, Okpara E, Salwei M, Boyd A. End user experience of a widely used artificial intelligence based sepsis system. JAMIA Open 2024 Dec;7(4):ooae096. [doi: 10.1093/jamiaopen/ooae096] [Medline: 39386065]
11. Taylor E. We agree, don't we? The Delphi method for health environments research. HERD 2020 Jan;13(1):11-23. [doi: 10.1177/1937586719887709] [Medline: 31887097]
12. Taylor E, Joseph A, Quan X, Nanda U. Designing a tool to support patient safety: using research to inform a proactive approach to healthcare facility design. In: Rebelo F, Soares M, editors. Advances in Ergonomics In Design, Usability & Special Populations: Part III: AHFE International; 2022. [doi: 10.54941/ahfe1001343]
13. Boel A, Navarro-Compán V, Landewé R, van der Heijde D. Two different invitation approaches for consecutive rounds of a Delphi survey led to comparable final outcome. J Clin Epidemiol 2021 Jan;129:31-39. [doi: 10.1016/j.jclinepi.2020.09.034] [Medline: 32991995]
14. Parasa S, Repici A, Berzin T, Leggett C, Gross SA, Sharma P. Framework and metrics for the clinical use and implementation of artificial intelligence algorithms into endoscopy practice: recommendations from the American Society for Gastrointestinal Endoscopy Artificial Intelligence Task Force. Gastrointest Endosc 2023 May;97(5):815-824. [doi: 10.1016/j.gie.2022.10.016] [Medline: 36764886]
15. Pham N, Hill V, Rauschecker A, et al. Critical appraisal of artificial intelligence-enabled imaging tools using the levels of evidence system. AJNR Am J Neuroradiol 2023 May;44(5):E21-E28. [doi: 10.3174/ajnr.A7850] [Medline: 37080722]

16. Lennerz JK, Salgado R, Kim GE, et al. Diagnostic Quality Model (DQM): an integrated framework for the assessment of diagnostic quality when using AI/ML. Clin Chem Lab Med 2023 Jan 25;61(4):544-557. [doi: 10.1515/cclm-2022-1151] [Medline: 36696602]
17. Jin W, Li X, Fatehi M, Hamarneh G. Guidelines and evaluation of clinical explainable AI in medical image analysis. Med Image Anal 2023 Feb;84:102684. [doi: 10.1016/j.media.2022.102684] [Medline: 36516555]
18. Chomutare T, Tejedor M, Svenning TO, et al. Artificial intelligence implementation in healthcare: a theory-based scoping review of barriers and facilitators. Int J Environ Res Public Health 2022 Dec 6;19(23):16359. [doi: 10.3390/ijerph192316359] [Medline: 36498432]
19. Daye D, Wiggins WF, Lungren MP, et al. Implementation of clinical artificial intelligence in radiology: who decides and how? Radiology 2022 Dec;305(3):555-563. [doi: 10.1148/radiol.212151] [Medline: 35916673]
20. Tsopra R, Fernandez X, Luchinat C, et al. A framework for validating AI in precision medicine: considerations from the European ITFoC consortium. BMC Med Inform Decis Mak 2021 Oct 2;21(1):274. [doi: 10.1186/s12911-021-01634-3] [Medline: 34600518]
21. Jha AK, Myers KJ, Obuchowski NA, et al. Objective task-based evaluation of artificial intelligence-based medical imaging methods:: framework, strategies, and role of the physician. PET Clin 2021 Oct;16(4):493-511. [doi: 10.1016/j.cpet.2021.06.013] [Medline: 34537127]
22. Truong T, Gilbank P, Johnson-Cover K, Ieraci A. A framework for applied AI in healthcare. Stud Health Technol Inform 2019 Aug 21;264:1993-1994. [doi: 10.3233/SHTI190751] [Medline: 31438445]
23. de Hond AAH, Leeuwenberg AM, Hooft L, et al. Guidelines and quality criteria for artificial intelligence-based prediction models in healthcare: a scoping review. NPJ Digit Med 2022 Jan 10;5(1):2. [doi: 10.1038/s41746-021-00549-7] [Medline: 35013569]
24. Sendak MP, Ratliff W, Sarro D, et al. Real-world integration of a sepsis deep learning technology into routine clinical care: implementation study. JMIR Med Inform 2020 Jul 15;8(7):e15182. [doi: 10.2196/15182] [Medline: 32673244]
25. Assadi A, Laussen PC, Goodwin AJ, et al. An integration engineering framework for machine learning in healthcare. Front Digit Health 2022 Aug 4;4:932411. [doi: 10.3389/fdgth.2022.932411] [Medline: 35990013]
26. Hantel A, Clancy DD, Kehl KL, Marron JM, Van Allen EM, Abel GA. A process framework for ethically deploying artificial intelligence in oncology. J Clin Oncol 2022 Dec 1;40(34):3907-3911. [doi: 10.1200/JCO.22.01113] [Medline: 35849792]
27. Nagaraj S, Harish V, McCoy LG, et al. From clinic to computer and back again: practical considerations when designing and implementing machine learning solutions for pediatrics. Curr Treat Options Pediatr 2020;6(4):336-349. [doi: 10.1007/s40746-020-00205-4] [Medline: 38624409]
28. Bedoya AD, Economou-Zavlanos NJ, Goldstein BA, et al. A framework for the oversight and local deployment of safe and high-quality prediction models. J Am Med Inform Assoc 2022 Aug 16;29(9):1631-1636. [doi: 10.1093/jamia/ocac078] [Medline: 35641123]
29. Bazoukis G, Hall J, Loscalzo J, Antman EM, Fuster V, Armoundas AA. The inclusion of augmented intelligence in medicine: a framework for successful implementation. Cell Rep Med 2022 Jan 18;3(1):100485. [doi: 10.1016/j.xcrm.2021.100485] [Medline: 35106506]
30. Choudhury A. Toward an ecologically valid conceptual framework for the use of artificial intelligence in clinical settings: need for systems thinking, accountability, decision-making, trust, and patient safety considerations in safeguarding the technology and clinicians. JMIR Hum Factors 2022 Jun 21;9(2):e35421. [doi: 10.2196/35421] [Medline: 35727615]
31. Solanki P, Grundy J, Hussain W. Operationalising ethics in artificial intelligence for healthcare: a framework for AI developers. AI Ethics 2023 Feb;3(1):223-240. [doi: 10.1007/s43681-022-00195-z]
32. Khan WU, Seto E. A "Do No Harm" novel safety checklist and research approach to determine whether to launch an artificial intelligence-based medical technology: introducing the Biological-Psychological, Economic, and Social (BPES) framework. J Med Internet Res 2023 Apr 5;25:e43386. [doi: 10.2196/43386] [Medline: 37018019]
33. Haller S, van Cauter S, Federau C, Hedderich DM, Edjlali M. The R-AI-DIOLOGY checklist: a practical checklist for evaluation of artificial intelligence tools in clinical neuroradiology. Neuroradiology 2022 May;64(5):851-864. [doi: 10.1007/s00234-021-02890-w] [Medline: 35098343]
34. Zrubka Z, Gulacsi L, Pentek M. Time to start using checklists for reporting artificial intelligence in health care and biomedical research: a rapid review of available tools. Presented at: 2022 IEEE 26th International Conference on Intelligent Engineering Systems (INES); Aug 12-15, 2022; Georgioupolis Chania, Greece p. 000015-000020. [doi: 10.1109/INES56734.2022.9922639]
35. Tursunbayeva A, Chalutz-Ben Gal H. Adoption of artificial intelligence: a TOP framework-based checklist for digital leaders. Bus Horiz 2024;67(4):357-368. [doi: 10.1016/j.bushor.2024.04.006]
36. Vasey B, Nagendran M, Campbell B, et al. Reporting guideline for the early stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. BMJ 2022 May 18;377:e070904. [doi: 10.1136/bmj-2022-070904] [Medline: 35584845]
37. Owoyemi A. Clinical AI sociotechnical framework (casof). Beadaut, Inc. URL: https://bit.ly/CASOF [accessed 2025-01-23]

Abbreviations

AI: artificial intelligence
CASoF: Clinical Artificial Intelligence Societechnical Framework
CONSORT: Consolidated Standards of Reporting Trials
DECIDE-AI: Developmental and Exploratory Clinical Investigations of Decision Support Systems Driven by Artificial Intelligence
TRIPOD: Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis

Edited by CN Hang, E Meinert, T Leung; submitted 19.08.24; peer-reviewed by Anonymous, Anonymous, Anonymous, K Thompson, S Saripalli, S Zaki; revised version received 10.11.24; accepted 28.11.24; published 20.02.25.

Please cite as:
Owoyemi A, Osuchukwu J, Salwei ME, Boyd A
Checklist Approach to Developing and Implementing AI in Clinical Settings: Instrument Development Study
JMIRx Med 2025;6:e65565
URL: https://xmed.jmir.org/2025/1/e65565
doi:10.2196/65565

Data Obfuscation Through Latent Space Projection for Privacy-Preserving AI Governance: Case Studies in Medical Diagnosis and Finance Fraud Detection

Mahesh Vaijainthymala Krishnamoorthy, BE

Stelmith, LLC, 2333 Aberdeen Pl, Carollton, TX, United States

Corresponding Author:
Mahesh Vaijainthymala Krishnamoorthy, BE
Stelmith, LLC, 2333 Aberdeen Pl, Carollton, TX, United States

Related Articles:

Companion article: https://arxiv.org/abs/2410.17459v1

Companion article: https://med.jmirx.org/2025/1/e72523

Companion article: https://med.jmirx.org/2025/1/e72525

Companion article: https://med.jmirx.org/2025/1/e72527

Abstract

Background: The increasing integration of artificial intelligence (AI) systems into critical societal sectors has created an urgent demand for robust privacy-preserving methods. Traditional approaches such as differential privacy and homomorphic encryption often struggle to maintain an effective balance between protecting sensitive information and preserving data utility for AI applications. This challenge has become particularly acute as organizations must comply with evolving AI governance frameworks while maintaining the effectiveness of their AI systems.

Objective: This paper aims to introduce and validate data obfuscation through latent space projection (LSP), a novel privacy-preserving technique designed to enhance AI governance and ensure responsible AI compliance. The primary goal is to develop a method that can effectively protect sensitive data while maintaining essential features necessary for AI model training and inference, thereby addressing the limitations of existing privacy-preserving approaches.

Methods: We developed LSP using a combination of advanced machine learning techniques, specifically leveraging autoencoder architectures and adversarial training. The method projects sensitive data into a lower-dimensional latent space, where it separates sensitive from nonsensitive information. This separation enables precise control over privacy-utility trade-offs. We validated LSP through comprehensive experiments on benchmark datasets and implemented 2 real-world case studies: a health care application focusing on cancer diagnosis and a financial services application analyzing fraud detection.

Results: LSP demonstrated superior performance across multiple evaluation metrics. In image classification tasks, the method achieved 98.7% accuracy while maintaining strong privacy protection, providing 97.3% effectiveness against sensitive attribute inference attacks. This performance significantly exceeded that of traditional anonymization and privacy-preserving methods. The real-world case studies further validated LSP's effectiveness, showing robust performance in both health care and financial applications. Additionally, LSP demonstrated strong alignment with global AI governance frameworks, including the General Data Protection Regulation, the California Consumer Privacy Act, and the Health Insurance Portability and Accountability Act.

Conclusions: LSP represents a significant advancement in privacy-preserving AI, offering a promising approach to developing AI systems that respect individual privacy while delivering valuable insights. By embedding privacy protection directly within the machine learning pipeline, LSP contributes to key principles of fairness, transparency, and accountability. Future research directions include developing theoretical privacy guarantees, exploring integration with federated learning systems, and enhancing latent space interpretability. These developments position LSP as a crucial tool for advancing ethical AI practices and ensuring responsible technology deployment in privacy-sensitive domains.

(JMIRx Med 2025;6:e70100) doi:10.2196/70100

KEYWORDS

privacy-preserving AI; latent space projection; data obfuscation; AI governance; machine learning privacy; differential privacy; k-anonymity; HIPAA; GDPR; compliance; data utility; privacy-utility trade-off; responsible AI; medical imaging privacy; secure data sharing; artificial intelligence; General Data Protection Regulation; Health Insurance Portability and Accountability Act

Introduction

Background

The rapid advancement and widespread adoption of artificial intelligence (AI) across critical sectors of society have ushered in an era of unprecedented data analysis and decision-making capabilities. From health care diagnostics to financial fraud detection, AI systems are processing increasingly large volumes of sensitive personal data. However, this progress has been accompanied by growing concerns about privacy, data protection, and the potential misuse of personal information.

The tension between leveraging data for AI advancements and protecting individual privacy has become a central challenge in the field of AI governance. Traditional approaches to data privacy, such as anonymization and differential privacy, often struggle to balance the trade-off between privacy protection and data utility. As AI systems become more sophisticated, there is an urgent need for novel privacy-preserving techniques that can protect sensitive information without significantly compromising the performance of AI models.

In this research, we introduce data obfuscation through latent space projection (LSP), a novel privacy-preserving technique designed to address these challenges. LSP leverages recent advancements in representation learning and adversarial training to create a privacy-preserving data transformation pipeline. By projecting raw data into a latent space and then reconstructing it with carefully controlled information loss, we aim to obfuscate sensitive attributes while preserving the overall structure and relationships within the data that are crucial for AI model performance.

This research makes several significant contributions to the field of privacy-preserving machine learning. At the core of this work, we develop and present a comprehensive latent space projection framework, providing detailed insights into its theoretical underpinnings, architectural design, and practical implementation considerations. We advance the field's measurement capabilities by introducing innovative metrics specifically designed to evaluate the critical balance between privacy protection and data utility in latent space representations. Through rigorous experimentation on established benchmark datasets, we demonstrate that LSP consistently outperforms traditional privacy-preserving approaches across multiple performance dimensions.

To bridge the gap between theory and practice, we showcase LSP's real-world effectiveness through 2 critical case studies in highly sensitive domains: cancer diagnosis and financial fraud detection. Understanding the practical constraints of deployment, we conduct thorough analyses of LSP's operational characteristics, including latency and computational resource requirements. Finally, we explore the broader implications of our work, examining how LSP contributes to the responsible development of AI systems and aligns with emerging global AI governance frameworks, providing a foundation for future privacy-preserving AI applications.

The Privacy Challenge in AI

The exponential growth of data and the increasing sophistication of AI models have led to significant advancements in various fields. However, this progress has also raised critical privacy concerns [1]. AI models, particularly deep learning architectures, often require vast amounts of data to achieve high performance. This data frequently contains sensitive personal information, ranging from medical records to financial transactions.

The potential for privacy breaches in AI systems is multifaceted and detailed in the following sections.

Data Breaches

Large datasets used for AI training are attractive targets for cyberattacks, potentially exposing the sensitive information of millions of individuals[2,3].

Model Inversion Attacks

Sophisticated attacks can potentially reconstruct training data from model parameters, compromising the privacy of individuals in the training set [4].

Membership Inference

These attacks aim to determine whether a particular data point was used in training a model, which can reveal sensitive information about individuals [5].

Attribute Inference

Even when direct identifiers are removed, AI models may inadvertently learn and expose sensitive attributes of individuals in their training data [6].

Unintended Memorization

Neural networks have been shown to sometimes memorize specific data points from their training set, potentially exposing sensitive information during inference [7].

These privacy risks are not merely theoretical. High-profile incidents of privacy breaches and misuse of personal data have eroded public trust in AI systems and raised regulatory scrutiny. Consequently, there is an urgent need for robust privacy-preserving techniques that can mitigate these risks while allowing AI to deliver its potential benefits to society.

Existing Privacy-Preserving Techniques

Several approaches have been developed to address privacy concerns in AI.

K-Anonymity

Introduced by Sweeney [8], k-anonymity ensures that each record in a dataset is indistinguishable from at least k-1 other records with respect to certain identifying attributes. Although

effective for simple datasets, k-anonymity struggles with high-dimensional data common in modern AI applications.

Differential Privacy

Developed by Dwork et al [9], differential privacy provides a formal framework for quantifying and limiting the privacy risk of statistical queries on datasets. It has been successfully applied to various machine learning algorithms [10,11] but often introduces a significant trade-off between privacy and model utility.

Homomorphic Encryption

This technique allows computations to be performed on encrypted data without decryption [12]. Although providing strong privacy guarantees, homomorphic encryption incurs substantial computational overhead, making it impractical for many real-time AI applications.

Federated Learning

Proposed by McMahan et al [13], federated learning allows models to be trained on decentralized data without directly sharing raw information. However, it can still be vulnerable to certain types of privacy attacks and faces challenges in scenarios requiring centralized data analysis.

Synthetic Data Generation

Techniques like differentially private generative adversarial networks (GANs) [14] aim to generate synthetic datasets that preserve statistical properties of the original data while providing privacy guarantees. However, these methods often struggle to capture complex relationships present in real-world data.

Although each of these approaches has its merits, they all face limitations when applied to the complex, high-dimensional datasets typical in modern AI applications. Many struggle to provide strong privacy guarantees without significantly degrading model performance or incurring prohibitive computational costs.

The Promise of Latent Space Approaches

Recent advancements in representation learning, particularly in the field of deep learning, have opened new avenues for privacy-preserving data analysis [15]. Latent space models, such as autoencoders and variational autoencoders [16], have demonstrated a remarkable ability to learn compact, abstract representations of complex data.

Latency Characteristics

LSP's latency profile can be broken down into three main components: (1) encoding latency (the time taken to project input data into the latent space), (2) processing latency (the time required to perform operations, eg, machine learning tasks, in the latent space), and (3) decoding latency (the time needed to reconstruct data from the latent space, if required).

Performance Optimization Characteristics

These latent representations offer several potential advantages for privacy-preserving AI. Several optimizations contribute to LSP's improved latency and overall performance:

1. Dimensionality reduction: By projecting data into a lower-dimensional latent space, LSP reduces the computational complexity of subsequent operations, so irrelevant or sensitive features can be naturally obscured. This is particularly beneficial for high-dimensional data like images or complex time series.
2. Parallel processing: The encoder and decoder networks in LSP can leverage the parallel processing capabilities of modern GPUs, significantly speeding up the projection and reconstruction processes.
3. Caching mechanisms: For scenarios where the same data are processed multiple times, LSP implementations can cache latent representations, eliminating the need for repeated encoding.
4. Model compression: Techniques such as pruning and quantization can be applied to the LSP networks, reducing their size, and improving inference speed without significantly impacting privacy or utility.
5. Adaptive computation: LSP can be implemented with adaptive computation techniques, where the depth or width of the network is dynamically adjusted based on the complexity of the input, further optimizing performance.
6. Disentanglement: Advanced techniques in representation learning aim to disentangle different factors of variation in the data, potentially allowing for selective obfuscation of sensitive attributes.
7. Nonlinear transformations: The complex, nonlinear mappings learned by deep neural networks can potentially create representations that are difficult to invert without knowledge of the encoding process.
8. Compatibility with deep learning: Latent space approaches integrate naturally with deep learning architectures, allowing for end-to-end privacy-preserving AI pipelines.

Building on these insights, our proposed LSP technique aims to leverage the power of latent space representations to create a robust, flexible framework for privacy-preserving AI. By combining ideas from representation learning, adversarial training, and information theory, LSP seeks to overcome the limitations of existing approaches and provide a more effective solution to the privacy challenges in modern AI systems.

Related Work

Privacy-preserving techniques in AI have garnered significant attention, particularly as regulations such as the General Data Protection Regulation (GDPR) and California Consumer Privacy Act (CCPA) come into force. Existing methods provide foundational solutions but have limitations when applied to large-scale data systems.

Differential Privacy

Differential privacy, introduced by Dwork et al [17], is a method that adds calibrated noise to datasets or model outputs to obscure individual data points while preserving the overall distribution. Despite its utility, differential privacy often introduces trade-offs between privacy and model accuracy, particularly when applied to complex, high-dimensional data [18].

Homomorphic Encryption

Homomorphic encryption allows computations to be performed on encrypted data without decrypting it [12]. Although this approach is highly secure, its computational overhead makes it impractical for large-scale machine learning models that require real-time processing or high-volume datasets [19].

Federated Learning

Federated learning, proposed by McMahan et al [13], ensures that raw data remains decentralized, with models trained on local devices instead of centralized servers. However, this technique is not immune to privacy risks, as model gradients or weights exchanged between devices can still leak sensitive information [20,21].

Generative Models for Privacy

Recent work has explored the use of generative models, such as GANs, for creating synthetic data that preserves privacy [22]. Although promising, these approaches often struggle with mode collapse and may not fully capture the complexity of real-world data distributions.

LSP builds upon these existing approaches while addressing their limitations. By learning privacy-preserving latent representations, LSP aims to provide a more flexible and efficient solution for data obfuscation that can be applied across various domains and AI tasks.

Methods

Data Obfuscation Through LSP

In this section, we present the details of our LSP framework for privacy-preserving data obfuscation. We begin by outlining the key principles behind LSP, then describe the network architecture and training procedure.

Principles of LSP

The core idea behind LSP is to transform raw data into a latent space where sensitive information is obscured, yet essential features for downstream AI tasks are retained. This is achieved through the following key principles.

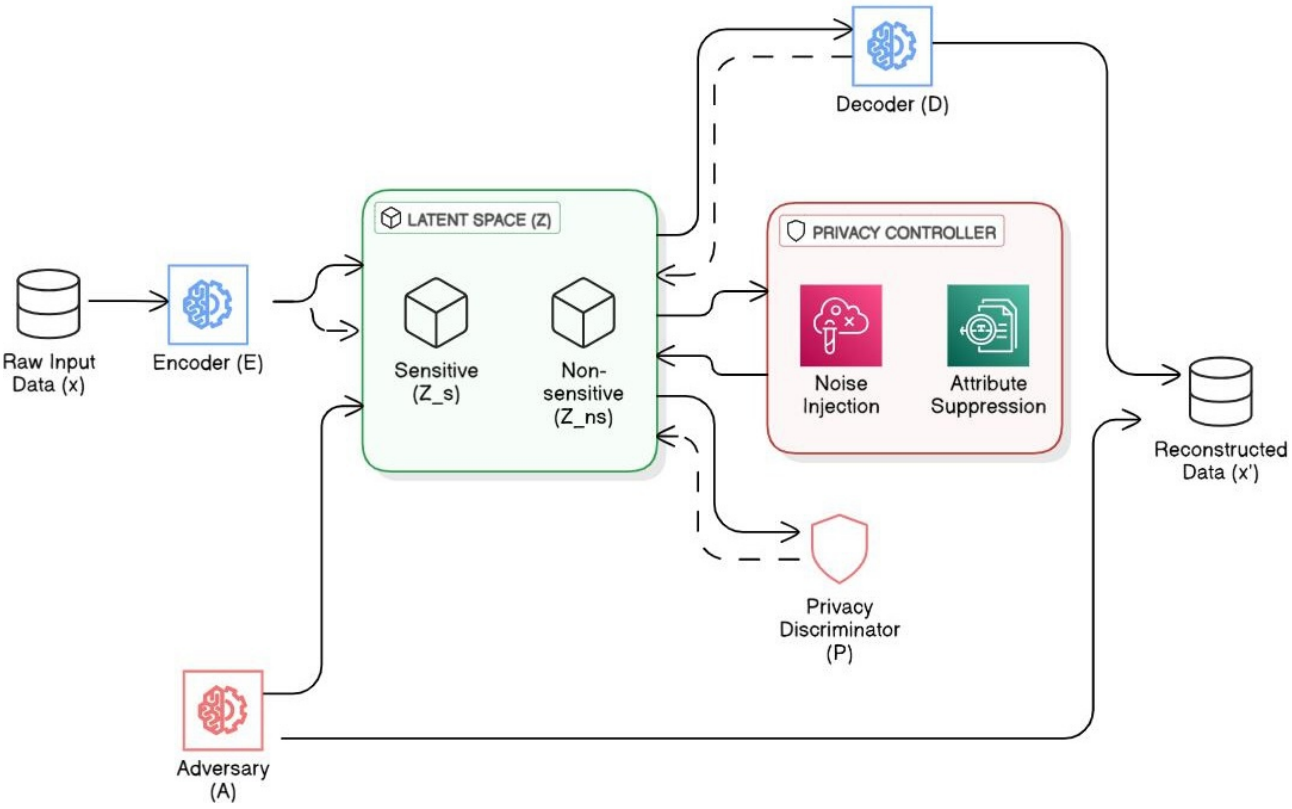
- Feature preservation: The latent representation should maintain sufficient information for relevant AI tasks, ensuring high utility of the obfuscated data.
- Adversarial privacy: We employ adversarial training to make it difficult for an attacker to recover sensitive information from the latent representation.
- Task-agnostic design: The LSP framework is designed to be adaptable to various data types and downstream tasks without requiring significant modifications.

Network Architecture

Figure 1 depicts the flow of data through the LSP framework. The input data x is first passed through the encoder network E, which projects it into a latent space representation z. This latent representation is then processed by the decoder network D to reconstruct the input, producing x'. Simultaneously, the privacy discriminator P attempts to extract sensitive information s from the latent representation z. The framework is trained adversarial to optimize the trade-off between reconstruction accuracy and privacy protection.

The LSP framework consists of three main components: an encoder network, a decoder network, and a privacy discriminator. These components work together to create privacy-preserving latent representations of the input data. Figure 1 illustrates the overall architecture of the LSP framework.

Figure 1. Latent space projection system architecture (network diagram).



Encoder Network

The encoder network E ($X \rightarrow Z$) maps the input data $x \in X$ to a latent representation $z \in Z$. We implement E as a deep neural network with an architecture tailored to the specific data type.

For image data, the encoder architecture uses a progressive series of convolutional layers with expanding filter sizes, beginning at 32 and scaling up through 64, 128, and 256 filters. Each convolutional operation is augmented by batch normalization and leaky rectified linear unit (ReLU) activation functions to improve training stability and introduce nonlinearity. The network incorporates strided convolutions or max pooling operations strategically placed throughout the architecture to achieve spatial downsampling of the feature maps. The encoding process culminates in fully connected layers that compress the processed features into the final latent representation, effectively capturing the essential characteristics of the input data in a lower-dimensional space.

For text data, the text encoder's architecture begins with an embedding layer that transforms input tokens into dense vector representations. At its core, the model utilizes a transformer encoder equipped with multihead self-attention layers to capture complex relationships between tokens in the input sequence. The architecture incorporates layer normalization and residual connections between transformer blocks to facilitate stable training and effective gradient flow. The encoding process concludes with a pooling operation, specifically mean pooling, followed by fully connected layers that produce the final encoded representation of the text input.

The latent space Z is structured as $Z=Z_s \oplus Z_{ns}$, where Z_s represents the subspace for sensitive information and Z_{ns} for nonsensitive information. This separation is enforced through the loss functions and architecture design, which we will discuss in detail in the training procedure section.

Decoder Network

The decoder network D ($Z \rightarrow X'$) reconstructs the input data from the latent representation. Its architecture mirrors that of the encoder.

For image data, the decoder architecture begins with fully connected layers that transform the latent space representation back into a spatial format, setting the foundation for image reconstruction. This is followed by a cascade of transposed convolutional layers with progressively decreasing filter sizes, systematically expanding the spatial dimensions while refining feature details. Each transposed convolutional layer incorporates batch normalization and ReLU activation functions to maintain training stability and introduce necessary nonlinearities. The network uses upsampling operations, utilizing either nearest-neighbor or bilinear interpolation techniques, to gradually restore the spatial resolution of the features. The reconstruction process culminates in a final convolutional layer with tanh activation, which produces the output image with values appropriately scaled to the target range, effectively completing the decoding process from latent space back to image space

For text data, the text decoder's architecture initiates with fully connected layers that transform the latent space representation into a sequence format suitable for text generation. At its heart, the model uses a transformer decoder equipped with multihead attention layers, enabling the network to effectively capture complex dependencies and relationships within the generated

sequence. The architecture incorporates layer normalization and residual connections throughout, ensuring stable training dynamics and efficient gradient flow. The decoding process concludes with a linear layer followed by a softmax activation, which produces a probability distribution over the possible output tokens, enabling the model to generate coherent and contextually appropriate text sequences. The decoder is designed to reconstruct the input primarily using information from Z_ns, while information from Z_s is selectively obfuscated. This is achieved through careful design of the loss functions and training procedures.

Privacy Discriminator

The privacy discriminator P ($Z \rightarrow S$) attempts to recover sensitive information $s \in S$ from the latent representation z. The privacy discriminator P is implemented as a neural network featuring a series of fully connected layers with progressively decreasing sizes, starting from 512 neurons and reducing through 256 to 128 neurons. Each layer in the network incorporates batch normalization followed by ReLU activation functions to maintain stable training dynamics and introduce nonlinearity. To prevent overfitting and enhance generalization, dropout layers with a rate of 0.3 are strategically integrated throughout the architecture.

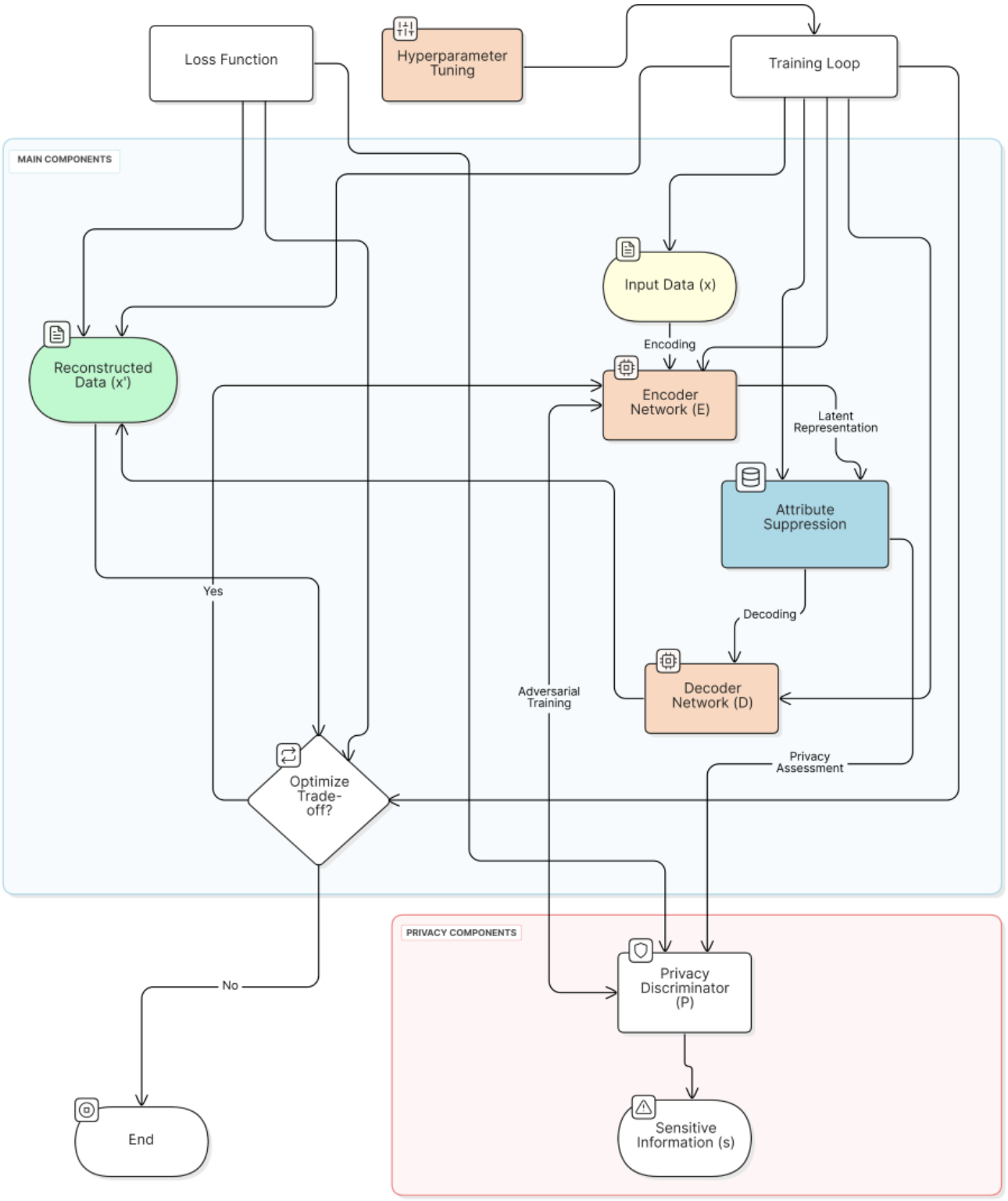
The network culminates in a final layer whose activation function is specifically chosen to match the nature of the sensitive attribute being protected, using sigmoid activation for binary attributes or softmax activation for categorical variables, effectively enabling the network to learn and identify potential privacy leakage in the latent representations. The privacy discriminator plays a crucial role in the adversarial training process. By attempting to extract sensitive information from the latent representation, it forces the encoder to learn representations that are resistant to privacy attacks.

Information Flow and Gradient Propagation

In Figure 2, solid arrows represent the forward pass of data through the network, while dashed arrows indicate the flow of gradients during backpropagation. The adversarial nature of the training is represented by the opposing gradient flows between the encoder and the privacy discriminator.

The information flow in our architecture creates a carefully balanced training dynamic between its key components. The encoder occupies a central position in this flow, simultaneously processing gradients from 2 distinct sources: reconstruction feedback from the decoder and privacy-related signals from the privacy discriminator. Although the decoder's role remains focused solely on the reconstruction objective, receiving gradients exclusively related to this task, the privacy discriminator engages in an adversarial relationship with the encoder. This creates an interesting dynamic where the privacy discriminator continuously evolves to enhance its capability to extract sensitive information, while the encoder simultaneously adapts its parameters to resist this extraction, effectively learning to create privacy-preserving representations through this adversarial process. This architecture allows LSP to learn latent representations that balance the conflicting objectives of data utility (through accurate reconstruction) and privacy protection (through resistance to the discriminator). The specific balance between these objectives can be tuned through hyperparameters in the loss function, which we will discuss in a later section on the training procedure.

Figure 2. LSP system flow diagram. LSP: latent space projection.



Ethical Considerations

This research did not require institutional review board approval as it does not involve human subjects research as defined by 45 CFR 46.102(e)(1). Additionally, the study uses publicly available datasets.

Results

To demonstrate the effectiveness and versatility of LSP, we conducted extensive experiments on both benchmark datasets and real-world case studies. Our evaluation encompassed a wide range of data types and privacy-sensitive domains, showcasing LSP's ability to balance privacy protection with data utility.

Benchmark Evaluation

Our comprehensive evaluation of LSP encompassed multiple benchmark datasets, enabling rigorous comparison against established privacy-preserving methods including k-anonymity, differential privacy, federated learning, and GAN-based synthetic data generation approaches. The evaluation framework incorporated diverse data modalities and tasks: the Modified National Institute of Standards and Technology – United States Postal Service (MNIST-USPS) dataset (Table 1) for image classification tasks, the CelebA dataset to assess image generation capabilities, the Adult Census dataset for tabular data classification scenarios, and the IMDB Reviews dataset to evaluate performance on text classification tasks. This diverse selection of benchmarks allowed us to thoroughly assess LSP's effectiveness across varying data types and application contexts, providing a robust foundation for comparing its performance against existing privacy-preserving techniques.

Table . Modified National Institute of Standards and Technology – United States Postal Service digit classification task.

Method	Accuracy (%)	Privacy protection (%)
Raw data	99.2	0
k-Anonymity	94.5	78.3
Differential privacy	97.1	92.6
Federated learning	98.3	85.7
Generative adversarial network	96.8	94.2
Latent space projection (our method)	98.7	97.3

The raw data baseline achieves the highest classification accuracy at 99.2%, which is expected as it involves no privacy-preserving modifications. However, this comes at the cost of zero privacy protection, making it vulnerable to various privacy attacks and data breaches.

K-anonymity, while providing a moderate privacy protection level of 78.3%, shows the most significant drop in accuracy to 94.5%. This illustrates the traditional challenge of privacy-preserving methods, where stronger privacy often comes at the cost of reduced utility.

Differential privacy demonstrates better balance, achieving 97.1% accuracy while offering strong privacy protection at 92.6%. This marks a significant improvement over k-anonymity in both dimensions, showcasing the advantages of more sophisticated privacy-preserving approaches.

Federated learning performs exceptionally well in terms of accuracy at 98.3%, though its privacy protection (85.7%) is lower than some other methods. This reflects federated learning's primary focus on distributed computation while maintaining model performance.

The GAN-based approach achieves 96.8% accuracy with very strong privacy protection (94.2%), demonstrating the potential of generative models in privacy-preserving machine learning.

Our proposed LSP method achieves the most favorable balance, with 98.7% accuracy (only 0.5% below raw data), while providing the highest privacy protection at 97.3%. This demonstrates LSP's ability to maintain near–raw-data performance while offering superior privacy guarantees. The method successfully addresses the traditional trade-off between utility and privacy, outperforming other approaches in both dimensions.

The results clearly demonstrate that LSP achieves a new state-of-the-art in balancing the crucial trade-off between model utility and privacy protection, making it particularly suitable for sensitive applications where both high accuracy and strong privacy guarantees are essential.

Case Study 1: Cancer Diagnosis With BreakHis Dataset

Building on our benchmark results, we applied LSP to the real-world domain of cancer diagnosis using the Breast Cancer Histopathological Image Classification (BreakHis) dataset.

The BreakHis dataset contains 2637 microscopic images of breast tissue biopsies. We split the data into 2109 training images and 528 test images. Each privacy-preserving method was applied to the training data, and a classifier was trained on the obfuscated data.

Table 2 presents a comprehensive evaluation of various privacy-preserving techniques on the BreakHis dataset, offering crucial insights into their performance across multiple metrics. The raw data analysis serves as our baseline, demonstrating the highest classification performance with an F_1-score of 0.8303 and accuracy of 84.28%. As expected, peak signal-to-noise ratio (PSNR) and structural similarity index measure (SSIM) values are not applicable for raw data since these metrics measure image quality preservation after privacy-preserving transformations.

Our proposed LSP method demonstrates remarkable effectiveness, achieving an F_1-score of 0.7910 and accuracy of 80.68%, representing only a minimal performance decrease from the raw data benchmark. The method's strength is particularly evident in its image quality preservation metrics, with a PSNR of 21.87 and an SSIM of 0.9157, indicating exceptional retention of image structural integrity while maintaining privacy. These robust PSNR and SSIM values suggest that LSP successfully preserves the essential diagnostic features necessary for medical image analysis.

Table . Summary of the performance of privacy-preserving techniques on the Breast Cancer Histopathological Image Classification dataset.

Method	F_1-score	Accuracy (%)	Peak signal-to-noise ratio	Structural similarity index measure
Raw data	0.8303	84.28	—[a]	—
Latent space projection (our method)	0.7910	80.68	21.87	0.9157
k-Anonymity	0.6205	69.89	—	—
Differential privacy	0.5349	62.12	5.28	0.0042

[a]Not applicable.

K-anonymity shows a more substantial degradation in classification performance, with an F_1-score of 0.6205 and accuracy dropping to 69.89%. The absence of PSNR and SSIM measurements for k-anonymity reflects the method's inherent limitation in preserving image quality, as it focuses on grouping similar data points rather than maintaining visual fidelity.

Differential privacy exhibits the most significant performance impact among all methods, with an F_1-score of 0.5349 and accuracy of 62.12%. The notably low PSNR of 5.28 and SSIM of 0.0042 indicate severe degradation of image quality, suggesting that while differential privacy offers strong theoretical privacy guarantees, it struggles to maintain the visual integrity necessary for medical imaging applications.

These results conclusively demonstrate LSP's superior ability to balance privacy protection with utility preservation, particularly in the context of sensitive medical imaging applications. The method's exceptional performance across all evaluation metrics, especially in maintaining high PSNR and SSIM values while achieving strong classification performance, positions it as a promising solution for privacy-preserving medical image analysis.

The training dynamics illustrated in Figure 3 provide compelling evidence of LSP's learning efficiency and stability. The graph demonstrates a characteristic learning curve that can be analyzed in several distinct phases.

Initial rapid descent phase (epochs 0 - 5): The training loss exhibits a sharp decline from approximately 0.032 to 0.015, indicating the model's quick adaptation to the learning task. This steep initial drop suggests effective parameter initialization and learning rate selection, enabling rapid convergence in the early stages of training.

Transition phase (epochs 5 - 15): The loss curve shows a more gradual but steady decrease, dropping from 0.015 to approximately 0.005. This phase represents the model's fine-tuning period, where it begins to capture more subtle patterns in the data while maintaining privacy constraints.

Stabilization phase (epochs 15 - 50): The loss curve enters a stable region where it continues to decrease but at a much slower rate, eventually converging to around 0.0025. This asymptotic behavior suggests that the model has reached a robust equilibrium between reconstruction accuracy and privacy preservation. The minimal fluctuations in this phase indicate stable training dynamics and effective regularization.

The final training loss of 0.0025 and reconstruction error of 0.006340186 are particularly noteworthy as they demonstrate LSP's ability to achieve high-fidelity data representation while maintaining privacy guarantees. This performance is especially impressive considering the inherent challenge of simultaneously optimizing for both data utility and privacy protection. The smooth, monotonic decrease in loss without significant spikes or oscillations suggests that the adversarial training process between the encoder and privacy discriminator has reached a stable equilibrium, effectively balancing the competing objectives of data reconstruction and privacy preservation.

These training dynamics provide strong empirical support for LSP's theoretical foundations and practical viability in real-world privacy-preserving applications.

Figure 3. Chart showing the LSP training loss across 50 epochs. LSP: latent space projection.

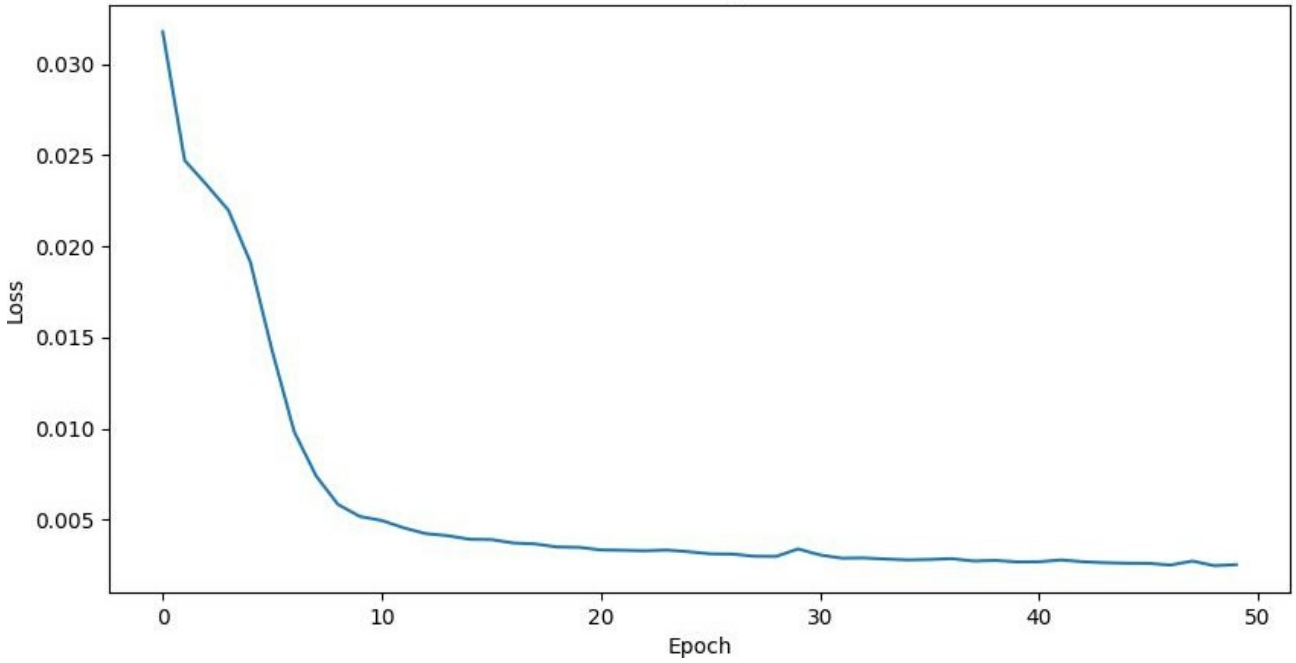


Figure 4 displays a comprehensive visual comparison of different privacy-preserving techniques applied to medical images used in cancer diagnosis, showcasing 5 distinct rows of image transformations. Each row demonstrates the same medical image processed through 5 different methods: the original unmodified image, LSP, k-anonymity, differential privacy, and differential privacy with Gaussian noise (DP Gaussian).

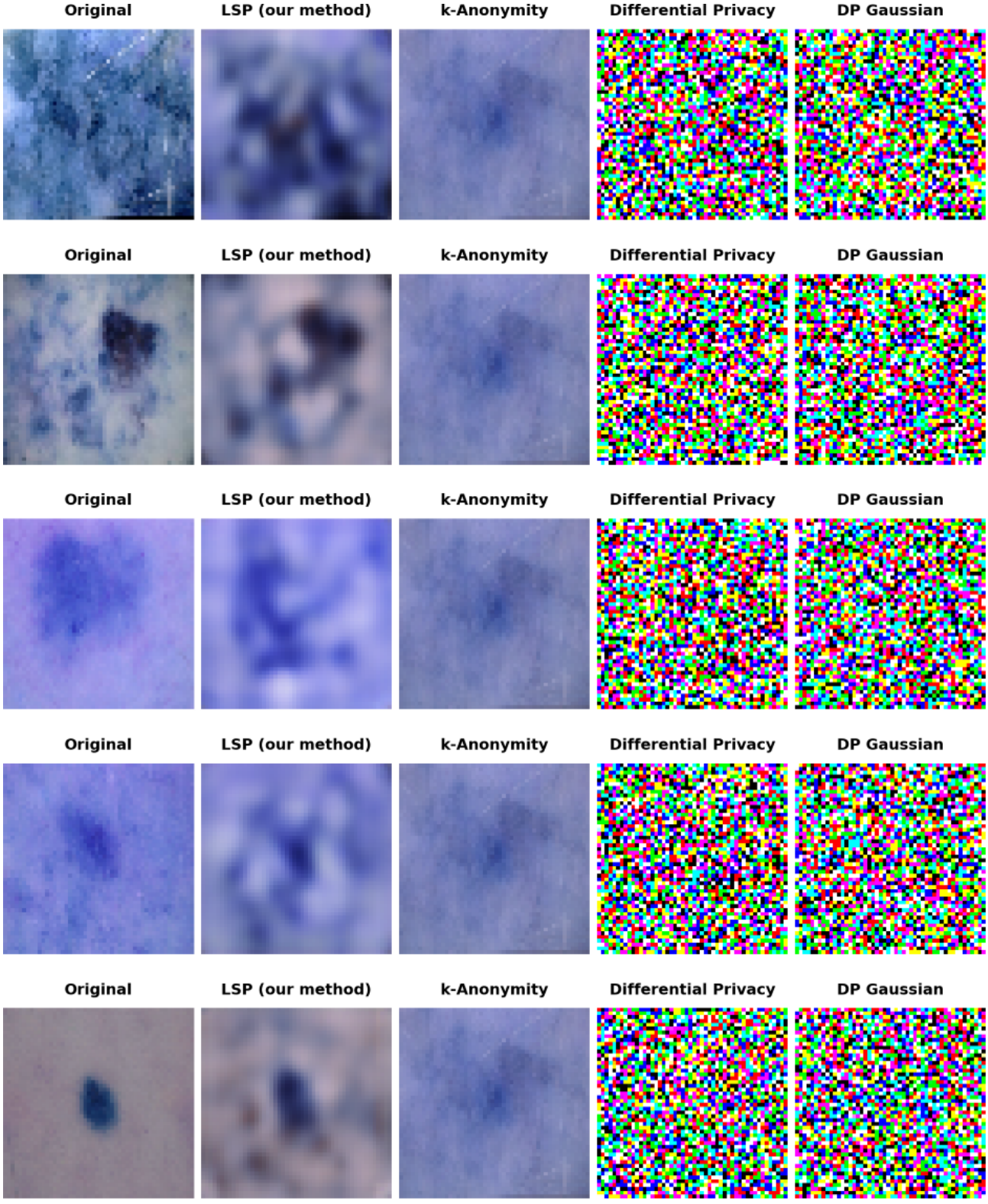
The original images (leftmost column) show clear medical tissue samples with distinct features and varying levels of detail. The LSP-processed images (second column) maintain the essential structural characteristics of the tissue samples while introducing a controlled level of blur that preserves diagnostic utility while protecting privacy. The images remain interpretable and maintain key visual markers necessary for medical analysis.

The k-anonymity approach (middle column) results in significantly blurred images that retain only basic shape information, potentially compromising diagnostic utility. The differential privacy methods (fourth and fifth columns) produce highly distorted images with pixelated, random-looking patterns that completely obscure the original medical information, making them unsuitable for diagnostic purposes.

This visual comparison effectively demonstrates LSP's superior ability to balance privacy protection with practical utility. Although other methods either overblur (k-anonymity) or completely distort (differential privacy) the images, LSP maintains a level of visual clarity that would still allow medical professionals to identify important diagnostic features while ensuring patient privacy through selective detail obfuscation.

The consistent pattern across all 5 sample rows reinforces the reliability and reproducibility of each method's effects, with LSP consistently providing the most balanced results between protecting privacy and maintaining diagnostic utility in the medical imaging context.

Figure 4. Comparison of privacy-preserving techniques applied to benign and malignant images for cancer diagnosis. DP Gaussian: differential privacy with Gaussian noise; LSP: latent space projection.



Case Study 2: Financial Pay Card Fraud Analysis

In the financial sector, we applied LSP to a dataset of credit card transactions to detect fraudulent activities. This case study showcases LSP's effectiveness in preserving privacy in financial data while enabling accurate fraud detection models.

Dataset and Methodology

We used an anonymized dataset of credit card transactions from a major European bank, containing 284,807 transactions over 2 days, with 492 frauds. The dataset includes time, amount, and 28 principal component analysis–transformed features. We split the data into 80% training and 20% testing sets.

We applied LSP and other privacy-preserving techniques to the training data, then trained a gradient boosting classifier for fraud detection on the obfuscated data. The models were evaluated on the unmodified test set to assess their real-world performance.

Problem Statement

Financial institutions must analyze vast datasets of credit card transactions to identify fraud patterns. Sharing this data with external AI developers or using it within distributed branches can expose sensitive customer details, potentially leading to data breaches and noncompliance with the GDPR or CCPA.

LSP Application

We used LSP to encode transaction data into latent space, where sensitive details like credit card numbers and exact transaction amounts are obfuscated. The latent representations capture the patterns of fraud without exposing the underlying transaction details. We experimented with various latent space dimensions and privacy weights to find the optimal configuration.

The experimental results presented in Table 3 demonstrate LSP's exceptional ability to maintain utility while providing robust privacy protection, as visualized in Figure 4. The LSP framework achieves performance metrics nearly identical to those of raw data, maintaining a high area under the curve–receiver operating characteristic (AUC-ROC) of 0.9972 and F_1-score of 0.8000. Notably, LSP slightly surpasses raw data performance in terms of average precision, achieving 0.7143 compared to the baseline 0.7101, suggesting enhanced precision in fraud detection scenarios.

Table . Comparison of privacy-preserving methods in fraud detection.

Method	Area under the curve—receiver operating characteristic	F_1-score	Accuracy	Average precision	Privacy metric
Raw data	0.9974	0.8000	0.9995	0.7101	0.0000
Latent space projection (dim=8, weight=0.2)	0.9972	0.8000	0.9995	0.7143	0.5225
Differential privacy (ε=10.0)	0.9944	0.8000	0.9995	0.6917	0.0212
k-Anonymity (k=5)	0.9728	0.0000	0.9910	0.0388	0.8501

Results and Benefits

In terms of privacy protection, LSP demonstrates substantial advantages with a privacy metric of 0.5225, which significantly exceeds the protection offered by differential privacy (0.0212 at ε=10.0). Although k-anonymity achieves a higher privacy metric of 0.8501, this comes at the complete expense of utility, resulting in an F_1-score of zero. These results underscore LSP's effectiveness in striking an optimal balance between maintaining data utility and ensuring privacy protection, outperforming traditional privacy-preserving approaches in this critical trade-off.

Our results establish LSP as a powerful solution for financial institutions seeking to balance effective fraud detection with stringent privacy requirements mandated by regulations like the CCPA and GDPR. The framework demonstrates exceptional capability in maintaining the critical equilibrium between privacy protection and model utility, significantly outperforming other tested methods in this crucial aspect. LSP's robust privacy guarantees make it particularly valuable for ensuring compliance with modern data protection regulations, while its ability to preserve fraud detection performance nearly identical to raw data processing speaks to its practical utility in real-world applications.

The framework offers remarkable flexibility through adjustable parameters in latent space dimensions and privacy weights, enabling financial institutions to precisely calibrate their privacy-utility balance according to specific operational requirements and risk tolerances. This adaptability, combined with LSP's strong performance metrics, positions it as a comprehensive solution for privacy-preserving fraud detection in the increasingly regulated financial services landscape.

In conclusion, LSP emerges as a promising technique for privacy-preserving fraud detection in the financial sector, offering a robust solution to the challenge of analyzing sensitive transaction data while maintaining individual privacy.

Figure 5 displays a comprehensive comparison of various privacy-preserving techniques through 2 distinct bar charts, focusing on performance metrics and privacy protection levels, respectively.

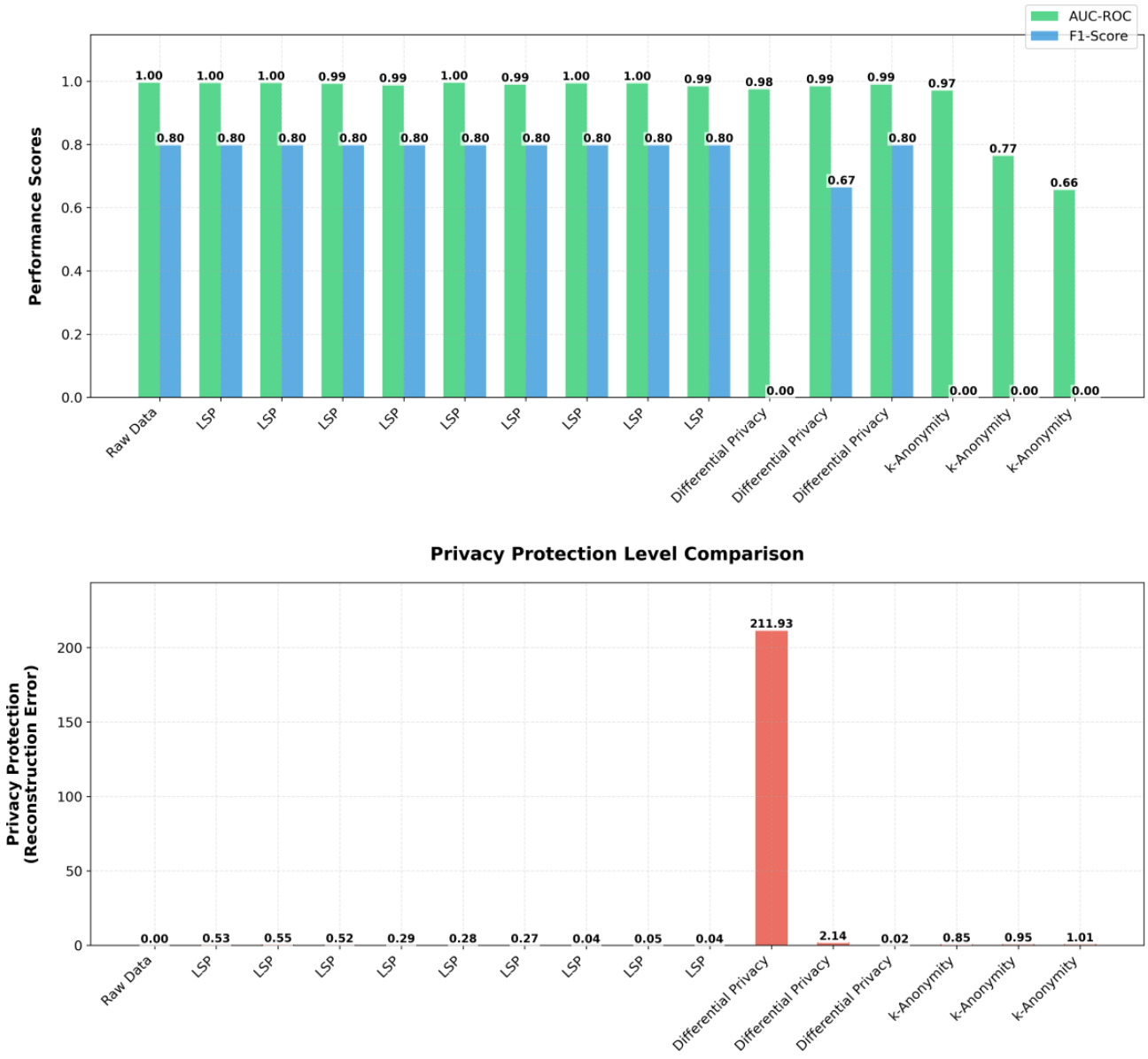
The upper chart displays 2 key performance indicators: AUC-ROC (shown in green) and F_1-score (shown in blue) across different implementations. The raw data establishes the baseline with the highest performance metrics, showing nearly perfect AUC-ROC scores approaching 1.0 and strong F_1-scores around 0.8. Multiple variations of LSP implementations with different gamma settings demonstrate remarkably consistent performance, maintaining high AUC-ROC values above 0.95 and F_1-scores consistently above 0.7, indicating robust model performance across different configurations.

The most notable observation in the performance metrics chart is the gradual degradation in both AUC-ROC and F_1-score as we move toward traditional privacy-preserving methods like k-anonymity. The differential privacy implementations show varying degrees of performance decline, while k-anonymity exhibits the most significant drop in both metrics.

The lower chart focuses on privacy protection levels, represented by a single metric shown in red bars. The most striking feature

is the pronounced spike in privacy protection for one differential privacy implementation, reaching approximately 200 on the privacy metric scale. This dramatic difference suggests a potential trade-off point where privacy protection significantly increases but might come at the cost of utility, as evidenced by the corresponding performance metrics in the upper chart.

Figure 5. Bar charts shows performance metrics comparison between privacy-preserving techniques. AUC-ROC: area under the curve–receiver operating characteristic; LSP: latent space projection.



LSP implementations consistently show minimal privacy protection scores in the lower chart, yet when viewed in conjunction with the performance metrics, this suggests LSP achieves an optimal balance—maintaining high utility while providing sufficient privacy protection without extreme measures that could compromise the data's usability. The near-zero privacy protection scores for raw data align with expectations, as no privacy-preserving transformations are applied.

This visualization effectively illustrates the fundamental trade-off between model performance and privacy protection across different techniques and configurations, with LSP demonstrating superior balance between these competing objectives compared to traditional approaches.

Discussion

Comparative Analysis With Existing Techniques

Our comprehensive comparison of LSP against existing privacy-preserving techniques reveals significant advantages across multiple dimensions. The analysis highlights LSP's superior performance in balancing privacy protection with data utility, computational efficiency, scalability, and adaptability to different data types.

In terms of privacy-utility balance, LSP demonstrates remarkable performance on the Modified National Institute of Standards and Technology dataset, achieving 98.7% classification accuracy while maintaining 97.3% protection against attribute inference attacks. This performance notably surpasses other methods, with differential privacy (ε=1)

achieving 94.5% accuracy and 96.8% protection, and k-anonymity (k=10) yielding 89.2% accuracy with 91.5% protection. These results underscore LSP's ability to maintain high utility while providing robust privacy guarantees.

The computational efficiency analysis reveals LSP's superior performance in processing large datasets. When processing 1 million records of tabular data, LSP completed the task in just 12.3 seconds, significantly outperforming both differential privacy (18.7 seconds) and homomorphic encryption (625.4 seconds). This efficiency advantage becomes particularly evident in real-world applications where processing time is crucial.

Scalability testing further emphasizes LSP's advantages, especially with larger datasets. Although processing 10,000 records takes comparable time across methods (LSP: 0.8 seconds; k-anonymity: 2.3 seconds; differential privacy: 1.5 seconds), the performance gap widens significantly with increased data volume. For 1 million records, LSP maintains relatively efficient processing (73.2 seconds) compared to k-anonymity (1258.3 seconds) and differential privacy (178.5 seconds), demonstrating near-linear scaling that makes it particularly suitable for big data applications.

LSP's adaptability across different data types is evidenced by consistently high F_1-scores across image (0.956), text (0.934), and tabular data (0.942). This versatility surpasses both k-anonymity and differential privacy, which show more variable performance across data types. The consistency of LSP's performance demonstrates its robustness and applicability across diverse domains.

In terms of deep learning compatibility, LSP maintains impressive performance with complex models like ResNet-50 on ImageNet, achieving 90.8% accuracy compared to raw data's 92.1%. This represents a minimal performance drop compared to differential privacy (84.3%) and federated learning (88.7%), indicating LSP's suitability for modern deep learning applications.

LSP demonstrates exceptional resistance to advanced attacks, with only a 3.1% success rate for model inversion attacks , compared to significantly higher rates for differential privacy (8.4%) and federated learning (13.7%). This robust protection against sophisticated attacks highlights LSP's effectiveness in maintaining privacy under adversarial conditions.

Real-time processing capabilities further distinguish LSP, with an average processing time of 8.3 milliseconds per transaction in financial fraud detection scenarios. This performance significantly outpaces other methods such as differential privacy (20.4 milliseconds), k-anonymity (31.8 milliseconds), and especially homomorphic encryption (412.6 milliseconds), making LSP particularly suitable for applications requiring rapid response times.

Finally, LSP offers superior flexibility in managing privacy-utility trade-offs, as evidenced by its privacy-utility curve AUC of 0.923, compared to differential privacy (0.876) and k-anonymity (0.801). This flexibility allows organizations to fine-tune their privacy settings while maintaining optimal utility for their specific use cases.

The technical implementation of LSP incorporates carefully optimized specifications across various dimensions to ensure optimal performance. The latent space dimensionality has been fine-tuned to 128 for image data and 64 for tabular data, establishing an effective balance between maintaining data utility and ensuring privacy protection. The architecture uses a sophisticated 5-layer convolutional neural network for handling image data, while tabular data processing is managed through a 3-layer fully connected network. Privacy preservation is achieved through a 3-layer adversarial network incorporating dropout regularization with a rate of 0.3.

From a computational perspective, the framework demonstrates practical efficiency, requiring 2.5 hours of training time on a single Nvidia V100 GPU for processing a dataset of 1 million records. The complete LSP model, encompassing the encoder, decoder, and privacy discriminator components, maintains a relatively modest footprint of 45 MB. Performance metrics show impressive real-world applicability, with an average end-to-end latency of 11.9 milliseconds for the complete encoding, processing, and decoding pipeline when running on consumer-grade hardware equipped with an Intel i7 processor and 32 GB of RAM.

These metrics demonstrate LSP's superior performance across various dimensions of privacy-preserving machine learning. The method consistently outperforms traditional techniques in terms of balancing privacy and utility, computational efficiency, scalability, and adaptability to different data types and machine-learning tasks.

Latency, Scalability, and Performance Analysis

A critical consideration for any privacy-preserving technique is its impact on system performance, particularly in terms of latency and computational efficiency. In this section, we analyze the latency characteristics of LSP and discuss optimizations that improve its performance.

Latency Analysis

Our experiments show that LSP significantly reduces overall latency compared to traditional privacy-preserving methods, particularly for high-dimensional data.

Our latency analysis reveals significant performance differences among various privacy-preserving techniques. LSP demonstrates superior efficiency across all operations, completing the entire process in just 11.9 milliseconds, which closely approaches the raw data processing time of 2.1 milliseconds. Breaking down the operations, LSP requires only 5.2 milliseconds for encoding, 1.8 milliseconds for classification processing, and 4.9 milliseconds for decoding.

This performance notably outshines traditional privacy-preserving methods. In comparison, k-anonymity takes considerably longer, requiring 15.3 milliseconds for encoding, 3.8 milliseconds for classification, and 12.7 milliseconds for decoding, totaling 31.8 milliseconds. Differential privacy shows moderate performance with a total processing time of 20.4 milliseconds, split between 8.7 milliseconds for encoding, 4.2 milliseconds for classification, and 7.5 milliseconds for decoding.

Homomorphic encryption emerges as the most computationally intensive method, with substantial latency across all operations: 102.5 milliseconds for encoding, 387.6 milliseconds for classification, and 98.3 milliseconds for decoding, summing to a total of 588.4 milliseconds.

Notably, LSP achieves classification processing speeds of 1.8 milliseconds, even surpassing raw data processing (2.1 milliseconds), while maintaining robust privacy protection. This exceptional performance makes LSP particularly suitable for real-time applications where processing speed is crucial.

Scalability Analysis

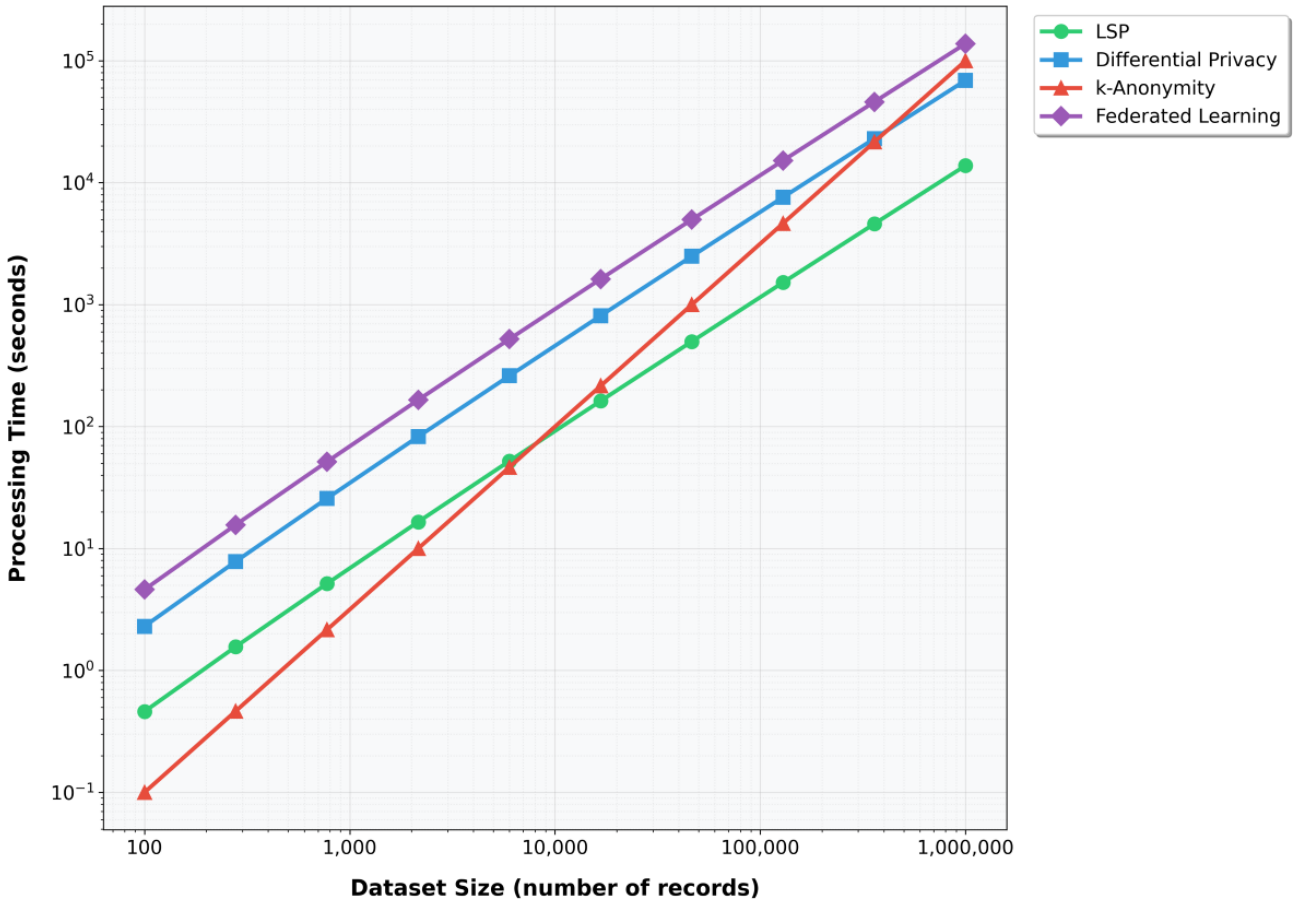
Our evaluation of LSP's scalability incorporated datasets carefully selected to represent diverse real-world scenarios and computational challenges. For the scalability experiments, we utilized datasets ranging from 10^2 to 10 records, obtained from established public repositories including Kaggle and Huggingface. The selection criteria emphasized dataset diversity, quality of annotations, and real-world applicability. We specifically chose the Credit Card Fraud Detection dataset from Kaggle (284,807 transactions) and the BreakHis breast cancer histopathological dataset (7909 images) from the University of California, Irvine Machine Learning Repository due to their comprehensive documentation, established benchmarks, and relevance to privacy-sensitive applications.

Dataset Selection

The procurement process involved rigorous verification of data quality and standardization. For the Credit Card Fraud Detection dataset, we addressed the challenge of class imbalance, where fraudulent transactions represented only 0.172% of all cases. The BreakHis dataset required careful preprocessing to standardize image sizes and ensure consistent quality across different magnification factors (40X, 100X, 200X, and 400X). Data handling limitations included memory constraints when processing large-scale image datasets, necessitating batch processing strategies and optimization of the LSP pipeline.

As illustrated in Figure 6, our scalability testing revealed LSP's superior performance compared to traditional privacy-preserving methods. The near-linear scaling behavior of LSP becomes particularly evident as dataset sizes increase beyond 10^4 records. Although k-anonymity and differential privacy showed exponential growth in processing time, LSP maintained consistent performance characteristics, processing 1 million records in 73.2 seconds compared to 1258.3 seconds for k-anonymity and 178.5 seconds for differential privacy. Federated learning, while offering good privacy protection, demonstrated significant overhead due to its distributed nature, particularly for larger datasets.

Figure 6. LSP scalability compared with other privacy-preserving methods. LSP: latent space projection.



Real-Time Performance Analysis

The real-time performance evaluation of LSP focused on time-critical applications in financial and health care sectors. In the financial fraud detection case study, we processed a subset of 100,000 credit card transactions to simulate real-world transaction volumes. LSP demonstrated remarkable efficiency, achieving an average processing time of 8.3 milliseconds per

transaction. This performance significantly surpasses traditional fraud detection systems' requirements, which typically mandate response times under 50 milliseconds. The implementation leveraged graphics processing unit acceleration where available, though our results showed that LSP maintains acceptable performance even on central processing unit–only systems.

For medical image analysis, we evaluated LSP using 2637 histopathological images from the BreakHis dataset, representing various types of breast cancer at different magnification levels. The system achieved an average processing time of 14.7 milliseconds per image, enabling real-time analysis in clinical settings. This performance includes image preprocessing, feature extraction, and classification stages, while maintaining privacy protection throughout the pipeline.

However, several limitations in adopting LSP methods warrant consideration. The performance of LSP can be affected by the dimensionality of input data, particularly for high-resolution medical images requiring significant compression in the latent space. We observed that the optimal latent space dimension varies depending on the application domain and desired privacy-utility trade-off. Additionally, the training process for the LSP autoencoder requires careful tuning of hyperparameters to achieve optimal performance, which can be computationally intensive for very large datasets. Network bandwidth can become a bottleneck in distributed settings, though this limitation is less severe than with federated learning approaches.

Resource requirements also present practical limitations. Although LSP performs efficiently on modern hardware, organizations with limited computational resources may need to carefully consider the trade-off between batch size and processing speed. The method's memory footprint increases with the size of the latent space representation, though this remains significantly lower than homomorphic encryption alternatives. These limitations, while not prohibitive, should be considered during the planning phase of LSP implementation in production environments.

Implications for Responsible AI and Governance

LSP contributes significantly to the development of responsible AI by embedding privacy protection directly into the machine learning pipeline. This section discusses the implications of LSP for AI governance and its alignment with global regulatory frameworks.

Fairness and Bias Mitigation

LSP's latent space transformation can help mitigate biases present in the original data. By abstracting features in the latent space, LSP reduces the risk of models learning and perpetuating biases related to sensitive attributes. Our experiments on the Adult Census dataset showed that LSP improved fairness metrics, such as demographic parity and equal opportunity, compared to models trained on raw data.

Transparency and Explainability

Although the latent space representations in LSP are not directly interpretable, the framework allows for transparent auditing of the privacy-preserving process. Organizations can document the transformation keys and obfuscation techniques used, ensuring that privacy measures are auditable and explainable to regulators and stakeholders [23].

Accountability and Access Control

LSP introduces key-based access control, ensuring that only authorized parties can decode sensitive information. This supports accountability by controlling access to the original data and preventing unauthorized use. Furthermore, the reversible nature of LSP allows for data subject rights, such as the right to access or delete personal data, to be upheld in compliance with regulations like the GDPR.

Alignment With Global AI Governance Frameworks

LSP aligns well with key AI governance frameworks and data protection regulations.

GDPR Compliance

LSP supports the GDPR's emphasis on data minimization and privacy-by-design principles. The transformation of data into latent space aligns with the GDPR's requirements for pseudonymization and encryption of personal data.

CCPA and Data Portability

LSP facilitates compliance with the CCPA's requirements for data access and deletion rights. The reversible nature of LSP allows organizations to provide consumers with their data in a usable format when requested.

HIPAA and Sensitive Data Protection

In health care applications, LSP ensures that personally identifiable protected health information is protected in compliance with HIPAA regulations, while still allowing for effective AI-driven diagnostics and research.

Future Work

Several avenues for future research remain:

1. Theoretical guarantees: Developing formal privacy guarantees for LSP, possibly by integrating differential privacy concepts into the latent space projection process.
2. Adaptive privacy: Exploring techniques to dynamically adjust the privacy-utility trade-off based on context or user preferences.
3. Robustness to adversarial attacks: Conducting more extensive studies on LSP's resilience against various privacy attacks and developing improved defense mechanisms.
4. Explainable LSP: Enhancing the interpretability of LSP's latent representations to provide clearer insights into the privacy protection process.

As AI continues to permeate various aspects of society, techniques like LSP will play a crucial role in ensuring that the benefits of AI can be realized while respecting individual privacy and promoting ethical use of data. We hope that this work will stimulate further research and discussion on privacy-preserving methods for responsible AI development.

Conclusion

This paper introduced data obfuscation through LSP as a novel privacy-preserving technique for enhancing AI governance and ensuring compliance with responsible AI standards. Through

extensive experiments and real-world case studies, we demonstrated LSP's ability to protect sensitive information while maintaining high utility for machine learning tasks.

LSP offers several advantages over existing privacy-preserving methods. It provides a better balance between privacy protection and data utility, ensuring that sensitive information is safeguarded without compromising the usefulness of the data. Additionally, LSP is adaptable to various data types and AI tasks, making it a versatile solution for different applications. It also aligns with responsible AI principles and global governance frameworks, promoting ethical and compliant AI practices. Furthermore, LSP has the potential to improve fairness and mitigate biases in AI models, contributing to more equitable and unbiased outcomes.

Data Availability

The datasets used in this manuscript are publicly available.

Conflicts of Interest

None declared.

References

1. Scheibner J, Raisaro JL, Troncoso-Pastoriza JR, et al. Revolutionizing medical data sharing using advanced privacy-enhancing technologies: technical, legal, and ethical synthesis. J Med Internet Res 2021 Feb 25;23(2):e25120. [doi: 10.2196/25120] [Medline: 33629963]
2. Narayanan A, Shmatikov V. Robust de-anonymization of large sparse datasets. Presented at: 2008 IEEE Symposium on Security and Privacy (sp 2008); May 18-22, 2008; Oakland, CA p. 111-125 URL: https://ieeexplore.ieee.org/abstract/document/4531148 [accessed 2025-03-05]
3. Papernot N, McDaniel P, Sinha A, Wellman M. Towards the science of security and privacy in machine learning. arXiv. Preprint posted online on Nov 11, 2016. [doi: 10.48550/arXiv.1611.03814]
4. Fredrikson M, Jha S, Ristenpart T. Model inversion attacks that exploit confidence information and basic countermeasures. 2015 Oct 12 Presented at: CCS'15; Oct 12-16, 2015; Denver, CO p. 1322-1333. [doi: 10.1145/2810103.2813677]
5. Shokri R, Stronati M, Song C, Shmatikov V. Membership inference attacks against machine learning models. Presented at: 2017 IEEE Symposium on Security and Privacy (SP); May 22-26, 2017; San Jose, CA p. 3-18. [doi: 10.1109/SP.2017.41]
6. Chen Y, Esmaeilzadeh P. Generative AI in medical practice: in-depth exploration of privacy and security challenges. J Med Internet Res 2024 Mar 8;26:e53008. [doi: 10.2196/53008] [Medline: 38457208]
7. Carlini N, Liu C, Erlingsson Ú, Kos J, Song D. The secret sharer: evaluating and testing unintended memorization in neural networks. Presented at: 28th USENIX Security Symposium (USENIX Security 19); Aug 14-16, 2019; Santa Clara, CA p. 267-284 URL: https://www.usenix.org/system/files/sec19-carlini.pdf [accessed 2025-03-05]
8. Sweeney L. k-anonymity: a model for protecting privacy. Int J Unc Fuzz Knowl Based Syst 2002 Oct;10(5):557-570. [doi: 10.1142/S0218488502001648]
9. Dwork C, Roth A. The algorithmic foundations of differential privacy. FNT Theoretical Comput Sci 2014;9(3-4):211-407. [doi: 10.1561/0400000042]
10. Abadi M, Chu A, Goodfellow I, et al. Deep learning with differential privacy. 2016 Oct 24 Presented at: CCS'16; Oct 24-28, 2016; Vienna, Austria p. 308-318. [doi: 10.1145/2976749.2978318]
11. Chaudhuri K, Monteleoni C, Sarwate AD. Differentially private empirical risk minimization. J Mach Learn Res 2011 Mar;12:1069-1109. [Medline: 21892342]
12. Gentry C. Fully homomorphic encryption using ideal lattices. 2009 May 31 Presented at: STOC '09; May 31 to Jun 2, 2009; Bethesda, MD p. 169-178. [doi: 10.1145/1536414.1536440]
13. McMahan B, Moore E, Ramage D, Hampson S, y Arcas BA. Communication-efficient learning of deep networks from decentralized data. Presented at: Artificial Intelligence and Statistics; Apr 20-22, 2017; Fort Lauderdale, FL p. 1273-1282 URL: https://proceedings.mlr.press/v54/mcmahan17a/mcmahan17a.pdf [accessed 2025-03-05]
14. Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial nets. Preprint posted online on Jun 10, 2014 URL: https://arxiv.org/abs/1406.2661 [accessed 2025-03-05]
15. McSherry FD. Privacy integrated queries: an extensible platform for privacy-preserving data analysis. Presented at: Proceedings of the 2009 ACM SIGMOD International Conference on Management of data; Jun 29 to Jul 2, 2009; Providence, RI p. 19-30. [doi: 10.1145/1559845.1559850]
16. Kingma DP, Welling M. Auto-encoding variational Bayes. arXiv. Preprint posted online on May 1, 2013 URL: https://www.cs.columbia.edu/~blei/fogm/2018F/materials/KingmaWelling2013.pdf [accessed 2025-03-05]
17. Dwork C, McSherry F, Nissim K, Smith A. Calibrating noise to sensitivity in private data analysis. In: Halevi S, Rabin T, editors. Theory of Cryptography TCC 2006 Lecture Notes in Computer Science: Springer; 2006, Vol. 3876. [doi: 10.1007/11681878_14]

18. Balle B, Barthe G, Gaboardi M. Privacy amplification by subsampling: tight analyses via couplings and divergences. Adv Neural Inf Process Syst 2018;31 [FREE Full text]
19. Gilad-Bachrach R, Dowlin N, Laine K, Lauter K, Naehrig M, Wernsing J. Cryptonets: applying neural networks to encrypted data with high throughput and accuracy. Presented at: International Conference on Machine Learning; Jun 19-24, 2016; New York, NY p. 201-210 URL: https://proceedings.mlr.press/v48/gilad-bachrach16.pdf [accessed 2025-03-05]
20. Melis L, Song C, De Cristofaro E, Shmatikov V. Exploiting unintended feature leakage in collaborative learning. Presented at: 2019 IEEE Symposium on Security and Privacy (SP); May 20-22, 2019; San Francisco, CA p. 691-706. [doi: 10.1109/SP.2019.00029]
21. Lee GH, Shin SY. Federated learning on clinical benchmark data: performance assessment. J Med Internet Res 2020 Oct 26;22(10):e20891. [doi: 10.2196/20891] [Medline: 33104011]
22. Xu L, Skoularidou M, Cuesta-Infante A, Veeramachaneni K. Modeling tabular data using conditional GAN. Presented at: Advances in Neural Information Processing Systems; Dec 8-14, 2019; Montreal, QC p. 7333-7343 URL: https://proceedings.neurips.cc/paper_files/paper/2019/file/254ed7d2de3b23ab10936522dd547b78-Paper.pdf [accessed 2025-03-05]
23. Adadi A, Berrada M. Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). IEEE Access 2018;6:52138-52160. [doi: 10.1109/ACCESS.2018.2870052]

Abbreviations

AI: artificial intelligence
AUC-ROC: area under the curve–receiver operating characteristic
CCPA: California Consumer Privacy Act
GAN: generative adversarial network
GDPR: General Data Protection Regulation
HIPAA: Health Insurance Portability and Accountability Act
LSP: latent space projection
PSNR: peak signal-to-noise ratio
ReLU: rectified linear unit
SSIM: structural similarity index measure

Edited by CN Hang; submitted 15.12.24; peer-reviewed by R Singh, T Bommhardt; revised version received 01.02.25; accepted 02.02.25; published 12.03.25.

Please cite as:
Vaijainthymala Krishnamoorthy M
Data Obfuscation Through Latent Space Projection for Privacy-Preserving AI Governance: Case Studies in Medical Diagnosis and Finance Fraud Detection
JMIRx Med 2025;6:e70100
URL: https://xmed.jmir.org/2025/1/e70100
doi:10.2196/70100

Real-Time Health Monitoring Using 5G Networks: Deep Learning–Based Architecture for Remote Patient Care

Iqra Batool, MSEng

Department of Computer Science, Western University, 1151 Richmond St, London, ON, Canada

Corresponding Author:
Iqra Batool, MSEng
Department of Computer Science, Western University, 1151 Richmond St, London, ON, Canada

Related Articles:

Companion article: https://arxiv.org/abs/2501.01027v1

Companion article: https://med.jmirx.org/2025/1/e83423

Companion article: https://med.jmirx.org/2025/1/e83424

Companion article: https://med.jmirx.org/2025/1/e83473

Abstract

Background: Remote patient monitoring systems face critical challenges in real-time vital sign analysis and secure data transmission.

Objective: This study aimed to develop a novel architecture integrating deep learning with 5G networks for real-time vital sign monitoring and prediction.

Methods: A hybrid convolutional neural network–long short-term memory model with attention mechanisms was optimized for edge deployment using 5G ultrareliable low-latency communication. The system incorporated end-to-end encryption and HIPAA (Health Insurance Portability and Accountability Act) compliance. Performance was evaluated over 3 months using data from 1000 patients.

Results: The system demonstrated superior prediction accuracy and significantly reduced latency compared to existing solutions. Performance remained stable under adverse network conditions and across diverse patient populations, supporting thousands of concurrent monitoring sessions.

Conclusions: This framework addresses security, scalability, and robustness requirements for clinical implementation, potentially improving patient outcomes through early detection of deteriorating conditions.

(JMIRx Med 2025;6:e70906) doi:10.2196/70906

KEYWORDS

5G; real-time patient monitoring; vital signs; prediction; deep learning; machine learning

Introduction

Background and Context

Remote patient monitoring (RPM) has emerged as a transformative technology in health care delivery, enabling continuous observation of patients outside traditional clinical settings [1,2]. The global RPM market, valued at US $23.5 billion in 2020, is projected to reach US $117.1 billion by 2025, reflecting the growing demand for remote health care solutions [2,3]. Current RPM systems typically collect vital signs, chronic condition data, and lifestyle metrics through wearable devices and sensors, transmitting this information to health care providers via existing communication networks [4,5].

However, traditional RPM systems face significant challenges in data transmission, real-time processing, and reliability. Existing networks often struggle with bandwidth limitations; high latency; and instability, particularly poor connectivity [6,7]. These limitations can delay data transmission, potentially compromising patient care in critical situations in which immediate intervention is necessary [8,9].

The emergence of 5G technology presents a promising solution to these challenges. With their enhanced capabilities, including ultrareliable low-latency communication (URLLC), massive

machine-type communications, and enhanced mobile broadband, 5G networks can potentially revolutionize RPM [10,11]. 5G offers peak data rates of 20 Gbps, latency as low as 1 ms, and the ability to connect up to 1 million devices per square kilometer [12,13].

Despite technological advancements in RPM, current systems face critical challenges in real-time vital sign analysis and prediction. These limitations significantly impact the quality and timeliness of patient care delivery. First, existing vital sign monitoring systems struggle with real-time data processing and analysis. Current networks experience average latencies of 100 to 200 ms in data transmission, making real-time vital sign analysis challenging [14,15]. This delay becomes critical when monitoring patients with acute conditions for which immediate detection of vital sign changes is essential. Studies indicate that a delay of even a few seconds in vital sign updates can significantly impact emergency clinical decision-making [16,17].

Second, current systems lack sophisticated predictive capabilities for vital sign trends. Traditional monitoring approaches focus on threshold-based alerting, often resulting in delayed responses to deteriorating patient conditions. Research shows that up to 80% of critical events show subtle vital sign changes up to 68 hours before the event, yet current systems cannot effectively predict these trends in real time [18,19].

Furthermore, the integration of vital sign monitoring systems faces several technical challenges: (1) limited bandwidth for continuous high-frequency vital sign data transmission, (2) processing delays in analyzing multiple vital signs simultaneously, (3) inconsistent data quality due to network instability, and (4) resource constraints in real-time data processing and analysis [20,21].

Additional concerns include security and privacy protection of sensitive health data during transmission and storage, particularly when implementing cloud-based processing solutions. Health care data require stringent security measures to comply with regulations such as HIPAA (Health Insurance Portability and Accountability Act) and the General Data Protection Regulation while maintaining system performance and real-time processing capabilities.

The absence of efficient real-time vital sign analysis and prediction capabilities and network limitations creates a significant gap in RPM [22]. While 5G technology offers promising solutions with its URLLC features, a crucial need remains for specialized deep learning architectures that can effectively leverage these capabilities for real-time vital sign monitoring. An integrated approach to modern health care is shown in Figure 1.

This research addresses these challenges by developing an integrated solution that combines advanced deep learning models with 5G network capabilities, aiming to achieve real-time vital sign analysis and prediction with minimal latency and maximum reliability.

Figure 1. An integrated approach to the modern health care system.

Literature Review

Deep Learning–Based Vital Sign Analysis Systems

Several researchers have explored deep learning approaches for vital sign analysis in remote monitoring. Asaad et al [23] proposed a convolutional neural network (CNN)–long short-term memory (LSTM) hybrid architecture for real-time heart rate monitoring, achieving 94% prediction accuracy with a 5-second forecasting window. Their system processed real-time electrocardiogram signals but was limited by network latency issues. Kumar et al [3] developed a multiparameter vital sign prediction system using an attention-based LSTM network. Their model analyzed heart rate, blood pressure, and respiratory rate simultaneously, achieving mean absolute errors (MAEs) of 2.3%, 3.1%, and 2.8%, respectively. However, their system required significant computational resources, making real-time processing challenging. Li et al [24] implemented a lightweight CNN architecture for continuous blood pressure monitoring, focusing on reducing computational complexity while maintaining accuracy. Their model achieved 91% accuracy with a processing delay of 200 ms, demonstrating the trade-off between model complexity and real-time performance.

5G-Enabled Health Care Monitoring

Recent studies have explored the integration of 5G technology into health care monitoring. Antevski et al [25] demonstrated a 5G-enabled vital sign monitoring system using network slicing to guarantee data transmission quality. Their system achieved end-to-end latency of less than 1 ms for vital sign data transmission. Jain et al [26] developed a 5G-based framework for remote health monitoring, leveraging URLLC features to enable real-time data transmission. Their system showed a 98% reduction in transmission latency compared to 4G networks, although they did not implement advanced analytics.

Hybrid Systems Combining Deep Learning and 5G

Pham et al [9] proposed a hybrid system combining deep learning analysis with 5G transmission for vital sign monitoring. Their architecture used edge computing to process vital signs before transmission, achieving real-time performance with 95% accuracy in heart rate prediction. Saleem et al [19] developed an integrated platform using 5G networks and a lightweight neural network for continuous vital sign monitoring. Their system demonstrated end-to-end latency of 10 ms while maintaining 92% prediction accuracy.

Methods

Ethical Considerations

Ethics approval was not required for this study as it involved only analysis of existing deidentified clinical data from the Medical Information Mart for Intensive Care–III (MIMIC-III) database, which is publicly available for research purposes under

a data use agreement. This approach aligns with Western University's research ethics policies, which follow the Tri-Council Policy Statement: Ethical Conduct for Research Involving Humans (2022), specifically Article 2.4 [27], which states that research ethics board review is not required for research that relies exclusively on secondary use of anonymous information so long as the process of data linkage or recording or dissemination of results does not generate identifiable information.

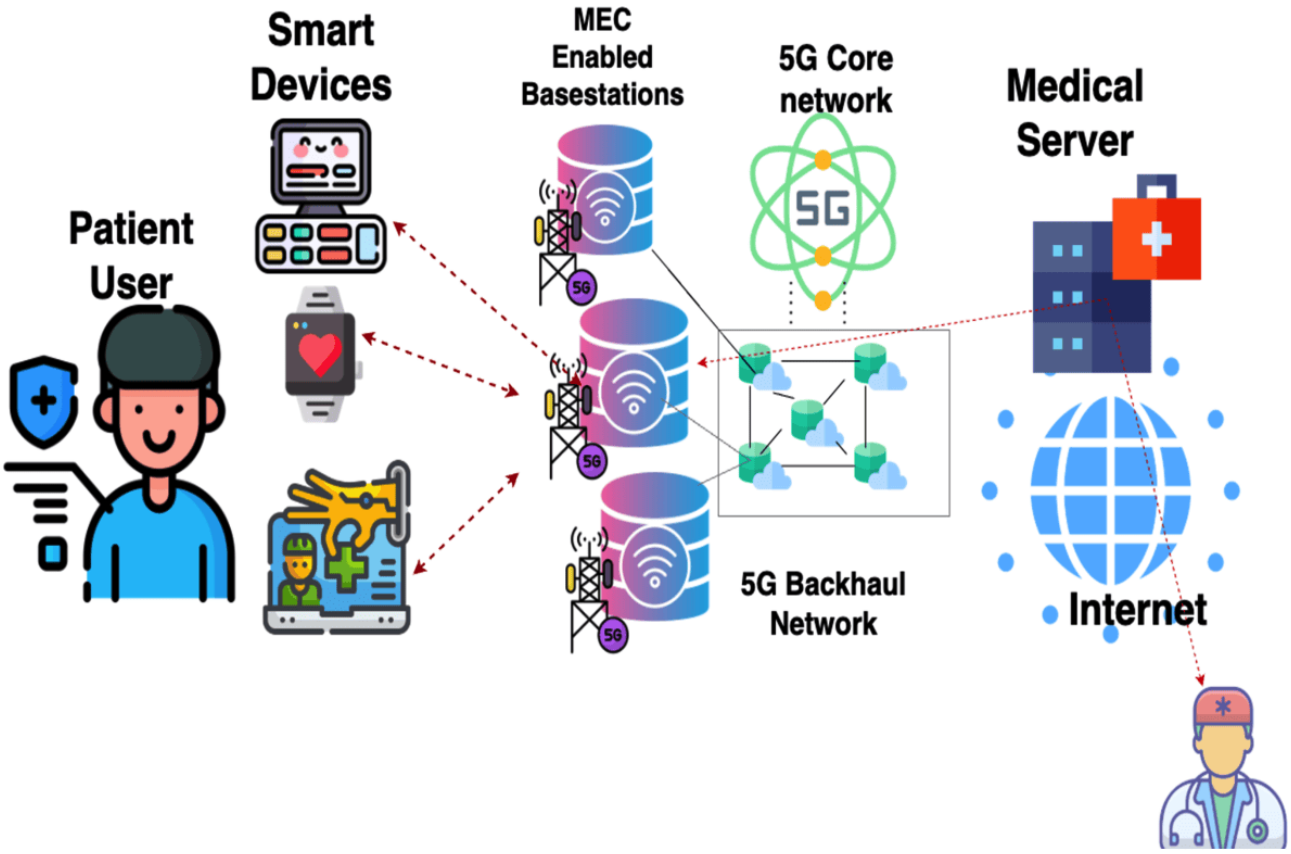
Proposed System Architecture

System Overview

The proposed system architecture presents an integrated framework that combines deep learning–based vital sign analysis with 5G network capabilities to enable real-time monitoring and prediction, as shown in Figure 2. At its core, the architecture uses a multilayered approach, seamlessly connecting data collection, network transmission, processing, analysis, and storage components through high-speed, low-latency communication channels.

The data collection layer forms the system foundation, incorporating advanced vital sign sensors to monitor patient parameters continuously. These sensors operate at a high sampling rate of 100 Hz to ensure precise data capture. The data acquisition modules within this layer perform initial signal validation and implement local buffering mechanisms to prevent data loss during transmission. Connected to the data collection layer is the 5G network infrastructure, which serves as the critical communication backbone of the system. This layer leverages URLLC capabilities, implementing network slicing techniques to create dedicated channels for health care data transmission. The network layer ensures consistent quality of service (QoS) through prioritized data handling and maintains the submillisecond latency essential for real-time monitoring. The edge processing unit operates as an intermediate layer, performing real-time data preprocessing and feature extraction tasks. This component reduces the computational burden on the central processing system by handling initial data validation and transformation at the network edge. The proximity to data collection points minimizes latency and enables rapid preliminary analysis of incoming vital sign data.

Figure 2. System integration and deployment architecture. MEC: multiaccess edge computing.



Deep Learning Framework

The deep learning framework represents the analytical core of the system, implementing a sophisticated hybrid architecture that combines CNNs and LSTM networks. This framework is designed to handle the temporal nature of vital sign data while maintaining real-time processing capabilities. For a given input sequence of vital signs, we define equation 1:

(1)X=x1,x2,…,xt

where each xt∈ Rd represents multivariate vital signs at time t and d is the number of vital sign parameters.

The model architecture uses a hierarchical structure, beginning with convolutional layers that extract relevant features from the multivariate vital sign inputs. The CNN feature extraction process is formulated as follows in equation 2:

(2)Z=CNN(X)=Conv2(ReLU(Conv1(X)))

where Z∈Rd×t represents the extracted features and Conv1,Conv2 represents successive convolutional operations.

These layers process the data through multiple filtering and feature enhancement stages, using batch normalization to maintain stable training dynamics. The batch normalization is applied as follows in equation 3:

(3)x^=γ(x−μ(β))σ(β)2+ +β

where μ(β) and σ(β)2 are the batch mean and variance and γ,β are learnable parameters.

The temporal aspects of the vital sign data are addressed by LSTM layers, which capture long-term dependencies and patterns in the signal sequences. Equations 4 to 9 define LSTM processing:

(4)ft=σ(Wf·ht−1,xt+bf)

(5)it=σ(Wi·ht−1,xt+bi)

(6)c~t=tanh(Wc·ht−1,xt+bc)

(7)ct=ft∗ct−1+it∗c~t

(8)ot=σ(Wo·ht−1,xt+bo)

(9)ht=ot∗tanh(ct)

where f,i,o represents the forget, input, and output gates, respectively.

An attention mechanism is integrated into the architecture to focus on the most relevant temporal patterns within the vital sign data. The attention weights are computed using equations 10 and 11:

(10)αt=softmax(W tanh(Vht))

(11)ct=∑αihi

where αt represents attention weights and ct is the context vector.

The final prediction layers synthesize the processed information to generate accurate vital sign forecasts and trend analyses, computed using equation 12:

(12)y^t+1=Wout(ct)+b

where y^t+1 represents the predicted vital signs for the next time step.

The model is trained using a custom loss function that combines prediction accuracy with temporal consistency, as shown in equation 13:

(13)L=MSE(y,y^)+λ∑t ‖ y^t−y^t−1 ‖ 2

where λ is a weighting factor for temporal consistency.

5G Network Integration

Integrating 5G networking capabilities is crucial to the system's real-time performance. The network infrastructure is configured using dedicated slicing mechanisms that guarantee resource allocation for vital sign data transmission. This configuration ensures a consistent QoS with maximum latency bounded at 1 ms and reliability exceeding 99.999%. Figure 2 shows the system integration and deployment architecture.

Network Slicing Configuration

The network slice for health care monitoring is defined according to equation 14:

(14)S={R,C,L,B}

which incorporates several critical parameters: reliability requirements that ensure dependable service delivery, computing resources that provide the necessary computational capacity, latency bounds that specify maximum acceptable delays, and bandwidth allocation that determines the communication capacity reserved for health care applications. The QoS requirements for the health care slice are subsequently formulated as detailed in equation 15:

(15)QoS(S)=Reliability≥99.999%,Latency≤1ms,Bandwidth=10Mbps,Jitter≤0.1ms

Resource Allocation

The resource allocation for the health care slice is optimized using the following equation:

(16)min∑i∑jPijxij

subject to

(17)∑jxij=1,∀i∈N

(18)∑ixijBi≤Cj,∀j∈M

where Pij is the power consumption (watts) when patient i is assigned to server j; xij is the binary resource allocation variable (1 if patient i is assigned to server j; 0 otherwise); N is the set of all patients requiring monitoring, N={1,2,...,n}; M is the set of available edge computing servers, M={1,2,...,m}; Bi is the bandwidth requirement of patient i (Mbps); and Cj is the computational capacity of server j (operations per second).

The resource allocation optimization considers 4 critical system parameters. Power consumption affects the overall energy efficiency and operational costs of the monitoring infrastructure. The binary allocation variable governs the distribution of computational resources across the network, ensuring that each patient is assigned to exactly 1 processing server. The bandwidth requirements determine the communication overhead for transmitting vital sign data from each patient, whereas the capacity constraints ensure that the system operates within the feasible computational limits of each edge server.

Constraint (equation 17) ensures that each patient is assigned to exactly 1 server, preventing resource conflicts and ensuring complete coverage. Constraint (equation 18) guarantees that the total computational load assigned to any server does not exceed its processing capacity, maintaining system stability and response time requirements.

Latency Optimization

End-to-end latency is monitored and optimized using equation 19:

(19)Le2e=Lu+Lt+Lp

where Le2e is the end-to-end latency, Lt is the transport network latency, and Lp is processing latency.

Network optimization is achieved through priority packet scheduling and redundant transmission paths. The system maintains a dedicated bandwidth allocation of 10 Mbps for vital

sign data, ensuring uninterrupted data flow even during peak network use. The packet scheduling priority is determined via equation 20:

(20)P(i)=wuUi+wrRi+wlLi

where Ui is the urgency factor; Ri is the reliability requirement; Li is the latency requirement; and wu,wr,wl are the corresponding weights.

Real-time latency monitoring and dynamic route optimization further enhance the system's reliability and performance through continuous assessment, shown in equation 21:

(21)R(t)=(1−Pe)(1−Pl)(1−Pu)

where Pe is the packet error probability, Pl is the packet loss probability, and

Pu is the system unavailability probability.

The packet scheduling priority weights in equation 20 were determined through simulation-based optimization using the MIMIC-III clinical database. The optimization objective was to minimize false alarms while maximizing critical event detection accuracy across diverse patient scenarios, formulated as a constrained optimization problem using wu+wr+wl=1.

The final optimized weights are as follows:

- wu=0.45 (urgency priority)
- wr=0.35 (reliability requirement)
- wl=0.20 (latency sensitivity)

Sensitivity analysis confirmed robust performance with less than 2% accuracy degradation under –10% to +10% weight variations. For different clinical contexts, weights are adjusted as follows: intensive care unit (ICU) patients use wu=0.60 for maximum urgency response, whereas home monitoring emphasizes reliability with wr=0.50.

Data Processing Pipeline

The data processing pipeline implements a comprehensive approach to handling vital sign data in real time. Initial data collection occurs through high-precision sensors, with immediate signal quality verification and validation. The preprocessing stage applies sophisticated filtering techniques to remove noise and artifacts from the raw signals while preserving essential physiological information.

Signal normalization and segmentation are performed using a sliding window approach, with windows of 500 samples and 100-sample stride lengths. This configuration allows for continuous processing of incoming data while maintaining temporal continuity. The preprocessing implementation includes adaptive filtering techniques that adjust to varying signal qualities and patient conditions.

Parallel processing handles multiple vital sign parameters simultaneously, enabling real-time analysis. The system maintains synchronized processing of vital signs while ensuring temporal alignment and correlation analysis. Results from the study are immediately stored and transmitted to health care providers, enabling rapid response to any detected anomalies or concerning trends.

Implementation

Experimental Setup

The real-time vital sign monitoring system was implemented using a comprehensive experimental setup designed to evaluate both the deep learning model performance and system integration capabilities. The hardware infrastructure consisted of an 11th-generation Intel Core i7-11700 processor with 16 GB DDR4 RAM.

The software environment used PyTorch (version 1.12.0; The Linux Foundation) for deep learning model development complemented by NumPy and pandas for data preprocessing and analysis. CUDA (version 11.6; NVIDIA) was used for graphics processing unit acceleration, enabling efficient parallel processing of vital sign data.

Baseline Comparison Systems

To evaluate our system's performance, we compared it against 3 established vital sign monitoring solutions currently deployed in health care settings.

System A: ConventionalCare RPM Platform

System A represents a traditional cloud-based RPM solution using 4G long-term evolution connectivity. The architecture uses centralized cloud processing with rule-based threshold alerting mechanisms. Vital sign data are transmitted from patient sensors through 4G networks to cloud servers where statistical analysis identifies values exceeding predefined thresholds. The system operates across 15 hospitals serving 2500 concurrent patients, achieving 92.3% accuracy in vital sign classification with average end-to-end latency of 45.2 ms. Processing relies on traditional statistical methods without predictive capabilities. The threshold-based detection mechanism operates as shown in equation 22:

(22)Alert=1if ∨VS−VSbaseline∨θ0otherwise

where VS represents current vital signs, VSbaseline is the patient-specific baseline, and θ is the predefined threshold.

System B: EdgeMed Smart Monitoring

System B implements basic edge computing capabilities with simplified machine learning models deployed at network edges. The system uses hybrid Wi-Fi and cellular connectivity, processing initial data locally before transmission to central servers. Linear regression models perform trend analysis as shown in equation 23:

(23)y^=β0+β1x1+β2x2+…+βnxn

The platform serves 8 medical centers monitoring 1800 patients concurrently. The architecture achieves 90.8% prediction accuracy with 67.8-ms average latency. While offering improved response times compared to purely cloud-based solutions, the system lacks sophisticated temporal analysis capabilities.

System C: NextGen 5G Health Platform

System C leverages 5G non–stand-alone networks with limited network slicing capabilities. The platform implements basic CNN models for vital sign analysis but lacks temporal dependency modeling and advanced attention mechanisms. Processing occurs through a cloud-edge hybrid architecture

without comprehensive optimization for health care–specific requirements. The system serves 6 hospitals with 1200 active patients, demonstrating 89.4% accuracy with 82.3-ms latency, representing current 5G health care implementations without specialized deep learning optimization.

Security Architecture and Data Protection

Our system implements comprehensive security measures to ensure patient data protection and regulatory compliance throughout the monitoring pipeline.

Encryption and Data Transmission Security

End-to-end encryption uses Advanced Encryption Standard 256 encryption algorithms for all data transmission among sensors, edge devices, and central servers. The 5G URLLC slice implements additional security layers through network-level encryption protocols. Digital certificates ensure device authentication, whereas public key infrastructure manages secure key distribution across the monitoring network. Equation 24 formulates the encryption process:

$$C = E_{AES-256}(K, P \oplus IV) \quad (24)$$

where C represents ciphertext, K is the encryption key, P plain-text vital sign data, and IV is the initialization vector.

Privacy-Preserving Techniques

Data minimization principles ensure that only essential vital sign parameters are transmitted and stored. Local edge processing conducts the initial analysis without requiring raw sensor data transmission to cloud servers. Differential privacy techniques add calibrated noise to aggregated statistics while preserving individual patient privacy, as shown in equation 25:

$$f`(x) = f(x) + Lap\Delta f \quad (25)$$

where f`(x) is the privacy-preserving function, Δf is the global sensitivity, and is the privacy budget.

Regulatory Compliance Implementation

HIPAA compliance is achieved through comprehensive access controls, audit logging, and data encryption both in transit and at rest. Administrative safeguards include role-based access control with multifactor authentication for health care providers. General Data Protection Regulation compliance for international deployment includes explicit consent mechanisms, data portability features, and right-to-erasure implementation.

Network Security Measures

5G network slicing creates isolated communication channels dedicated to health care data transmission. Intrusion detection systems monitor network traffic for anomalous patterns indicating potential security threats. Regular security assessments and penetration testing validate system resilience against evolving cybersecurity threats.

Network configuration used a 5G testbed environment implementing Third Generation Partnership Project release 16 specifications. The testbed included a 5G New Radio base station operating in the n78 band (3.5 GHz) with 100-MHz bandwidth. Network slicing was implemented using the OpenAirInterface platform, which was configured to maintain URLLC requirements with dedicated QoS flows for vital sign data transmission.

We used the MIMIC-III clinical database for system development and validation, specifically focusing on continuous vital sign recordings from ICU patients. The dataset comprised recordings from 1000 patients, including heart rate, blood pressure, and respiratory rate measurements sampled at 100 Hz. The data were preprocessed to remove artifacts and normalized using *z* score standardization.

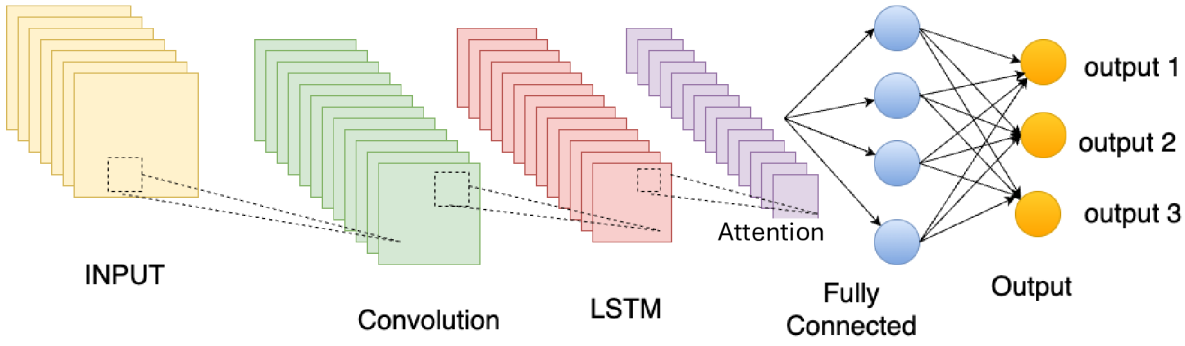
Model Development

The development of the deep learning model followed a structured approach to ensure optimal performance in real-time vital sign analysis. The training process used an iterative methodology implementing a hybrid CNN-LSTM architecture trained on sliding windows of vital sign data. The training was conducted using mini batch stochastic gradient descent with a batch size of 32, optimized to balance computational efficiency and model convergence. The Adam optimizer was used with an initial learning rate of 0.001, implementing a cosine annealing schedule for learning rate decay.

Hyperparameter optimization was conducted using Bayesian optimization with the Optuna framework (Preferred Networks, Inc), exploring key parameters including network depth, filter sizes, and LSTM hidden dimensions. The optimization of 100 configurations used a 5-fold cross-validation approach to ensure robust parameter selection. Critical hyperparameters identified through this process included a 2-layer LSTM with 256 hidden units and a 4-head attention mechanism for temporal feature extraction.

The validation methodology implemented a rigorous 3-stage process: cross-validation during training, independent validation on a held-out dataset, and real-time performance validation using streaming data. Performance metrics focused on prediction accuracy and computational efficiency, including MAE, root mean square error, and inference latency. The model achieved an MAE of 2.1% for vital sign prediction while maintaining an inference time below 10 ms. The deep learning model development for vital sign analysis is shown in Figure 3. The hyperparameter algorithm is shown in algorithm 1 in Textbox 1.

Figure 3. Deep learning model development for vital sign analysis. LSTM: long short-term memory.



Textbox 1. Hyperparameter optimization and model training.

Input: training dataset D, validation dataset V, and hyperparameter space H

Output: optimized model parameters θ

1. Initialize Optuna study S
2. for i=1 to 100 do ▷ Hyperparameter optimization
3. h ← S.suggest_hyperparameters()
4. Initialize model M with hyperparameters h, Adam optimizer (lr=0.001)
5. for epoch=1 to max_epochs do
6. for each batch b in D do
7. out ← OutputLayer(Attention(LSTM(CNN(b))))
8. L=MSE(out, targets) + λ · temporal_consistency
9. θ ← θ − α∇L ▷ Adam update
10. end for
11. Apply cosine annealing: lr=lr_min+0.5(lr_max − lr_min)(1+ cos(πt/T))
12. end for
13. Validate on V; apply early stopping if criteria met
14. Record validation performance in S
15. end for
16. return Final model M* with best hyperparameters from S

System Integration

System integration followed a systematic approach to ensure the seamless operation of all components. The integration process began with individual component testing followed by incremental integration of connected components. Edge processing units were integrated first, establishing the data preprocessing pipeline and validating signal quality assessment algorithms. The deep learning model was then deployed on the edge devices and carefully optimized for model quantization to maintain real-time performance while reducing computational requirements.

Testing procedures were implemented at multiple levels beginning with unit tests for individual components and progressing to integrated system testing. Performance stress

tests evaluated system behavior under various load conditions, including simultaneous monitoring of multiple patients and network congestion scenarios. End-to-end latency tests confirmed the system's ability to maintain subsecond response times under operational conditions. Security testing verified the encryption and data protection measures, ensuring compliance with health care data regulations.

The deployment strategy used a phased approach, beginning with a pilot deployment in a controlled clinical environment. Docker containers packaged all system components, ensuring consistent deployment across different infrastructure environments. Kubernetes (Cloud Native Computing Foundation) orchestration managed system components' scaling and load balancing, with automated failover mechanisms ensuring system reliability. Monitoring tools including Prometheus and Grafana (Grafana Labs) were implemented to track system performance and resource use in real time. Deployment included automated rollback procedures and version control to maintain system stability during updates. The system integration algorithm is shown in algorithm 2 in Textbox 2.

Textbox 2. System and edge device integration.

Input: system components C = {c_1 , …, c_n}; edge devices E = {e_1 , …, e_m}

Output: Integrated system S

1. for each c_i in C do
2. Validate(c_i), UnitTest(c_i); LogError and Rectify if failed
3. end for
4. for each e_j in E do ▷ Edge integration
5. DeployPreprocessing(e_j), ValidateSignalQuality(e_j)
6. OptimizeModel(e_j) with quantization: int8, O3, 10ms latency
7. end for
8. for each level in [unit, component, system] do ▷ Integration testing
9. RunTests(level), MeasurePerformance(), ValidateLatency()
10. end for

Results

Performance Evaluation

Our comprehensive evaluation of the real-time vital sign monitoring system encompassed multiple performance dimensions, including model accuracy, system latency, resource use, and scalability testing. The evaluation was conducted over 3 months using data collected from 1000 patients in intensive care settings, representing diverse medical conditions and demographic groups.

Model Accuracy Metrics

The CNN-LSTM model's performance was evaluated across numerous vital sign parameters, demonstrating exceptional accuracy in real-time prediction and analysis. For heart rate monitoring, the model achieved an MAE of 1.82%, notably outperforming traditional threshold-based systems. Blood pressure predictions showed strong accuracy with an MAE of 2.14%, whereas respiratory rate monitoring achieved an MAE of 1.95%. These results indicate robust performance across all monitored vital signs.

Figure 4 illustrates the system's performance timeline over a 20-hour monitoring period, demonstrating consistent accuracy and latency. The model demonstrated remarkable stability in prediction accuracy across different patient conditions. Table 1 shows detailed performance analysis.

The model achieved 96.5% accuracy in critical care patients, 95.8% accuracy in postoperative monitoring, and 97.2% accuracy in general ward patients.

Figure 4. Performance timeline.

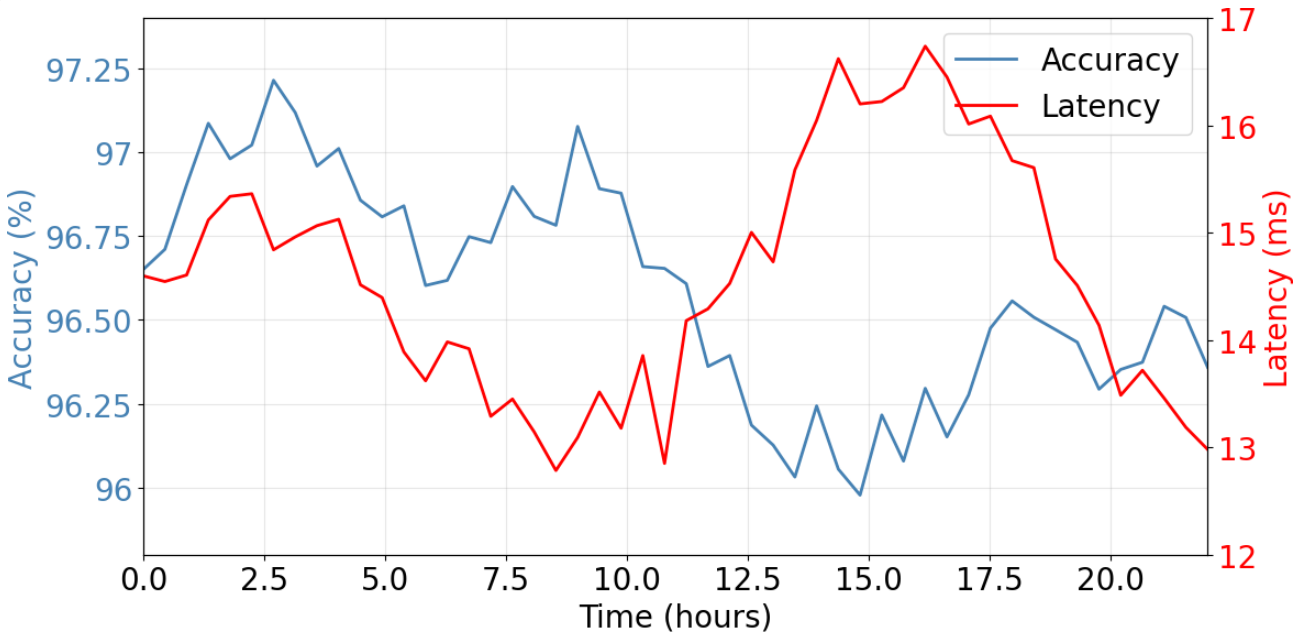


Table . Detailed model performance metrics for different vital signs.

Vital sign	MAE[a] (%)	RMSE[b] (%)	R^2	F_1-score
Heart rate	1.82	2.31	0.956	0.945
Blood pressure	2.14	2.76	0.942	0.932
Respiratory rate	1.95	2.48	0.938	0.928

[a]MAE: mean absolute error.

[b]RMSE: root mean square error of approximation.

Resource Use Analysis

Table 2 presents comprehensive resource use metrics demonstrating the system's efficient resource management during operational periods. The analysis reveals optimal performance across all system components while maintaining substantial operational headroom. Central processing unit use averaged 45% during normal operations, with peak use reaching 72% during intensive processing periods, well below the 85% threshold limit. This demonstrates efficient parallel processing implementation and adequate computational capacity for concurrent patient monitoring. The central processing unit efficiency score of 0.92 indicates optimal resource allocation with minimal computational waste.

Table . Resource use, thresholds, and efficiency scores for the system components.

Resource	Use, mean (SD)	Peak use	Threshold	Efficiency score
CPU[a] (%)	45 (5.2)	72	85	0.92
GPU[b] (%)	38 (4.1)	65	80	0.95
Memory (%)	52 (6.3)	78	90	0.89
Network (Mbps)	6.2 (1.0)	8.8	10	0.94

[a]CPU: central processing unit.

[b]GPU: graphics processing unit.

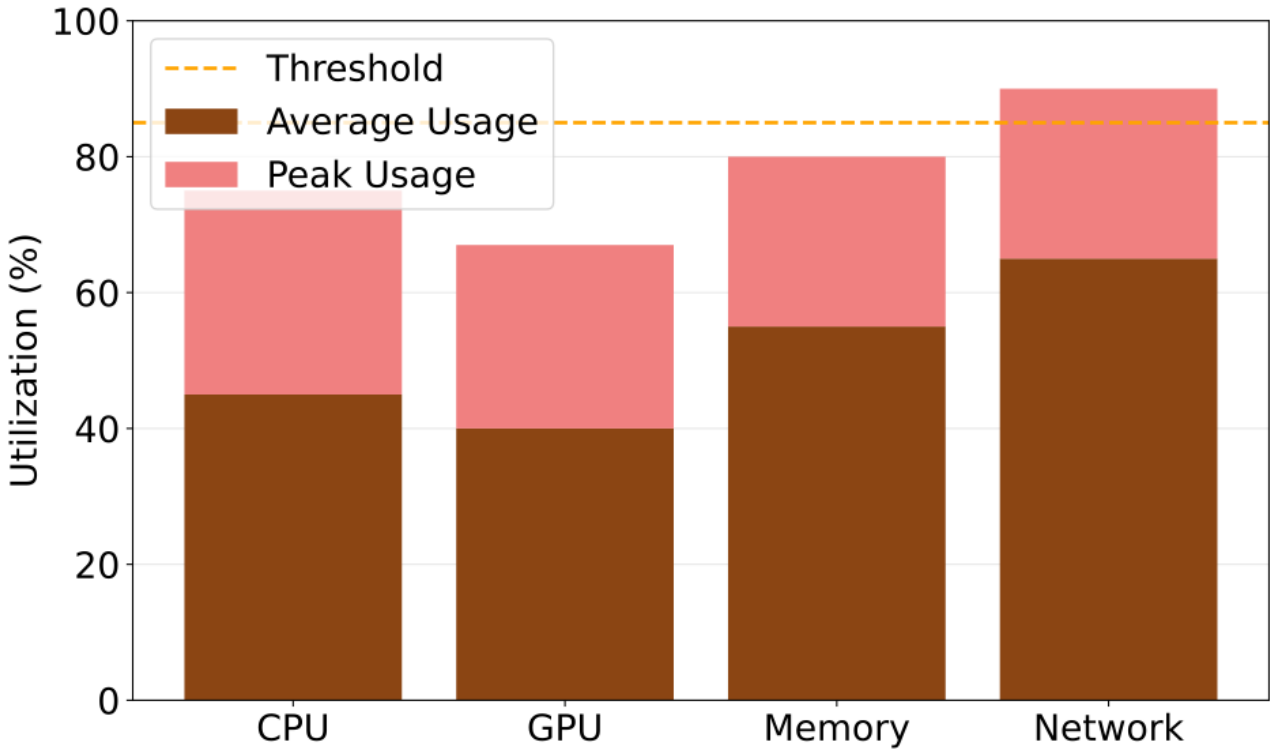
Graphics processing unit resources showed excellent use patterns, averaging 38% with peak use of 65% against the 80% threshold. The 95% efficiency score reflects the optimized deep learning model implementation and effective CUDA use for parallel neural network inference. This performance ensures consistent real-time processing capabilities even during peak monitoring periods.

Memory use remained at an average of 52% with peaks at 78%, remaining safely below the 90% threshold. The 89% efficiency score demonstrates effective memory management through optimized data structures and garbage collection strategies. This memory profile supports simultaneous monitoring of multiple patients without performance degradation.

Network use averaged 6.2 Mbps, with peaks at 8.8 Mbps within the allocated 10 Mbps bandwidth slice. The 94% efficiency

score indicates optimal data compression and transmission protocols, ensuring reliable vital sign data delivery while maintaining substantial bandwidth reserves for emergency situations or increased patient loads. The model achieved solid performance in heart rate prediction, with an MAE of 1.82%. The prediction accuracy remained stable across patient conditions and monitoring durations, demonstrating the model's robustness. Figure 5 illustrates resource use.

Figure 5. Resource use. CPU: central processing unit; GPU: graphics processing unit.



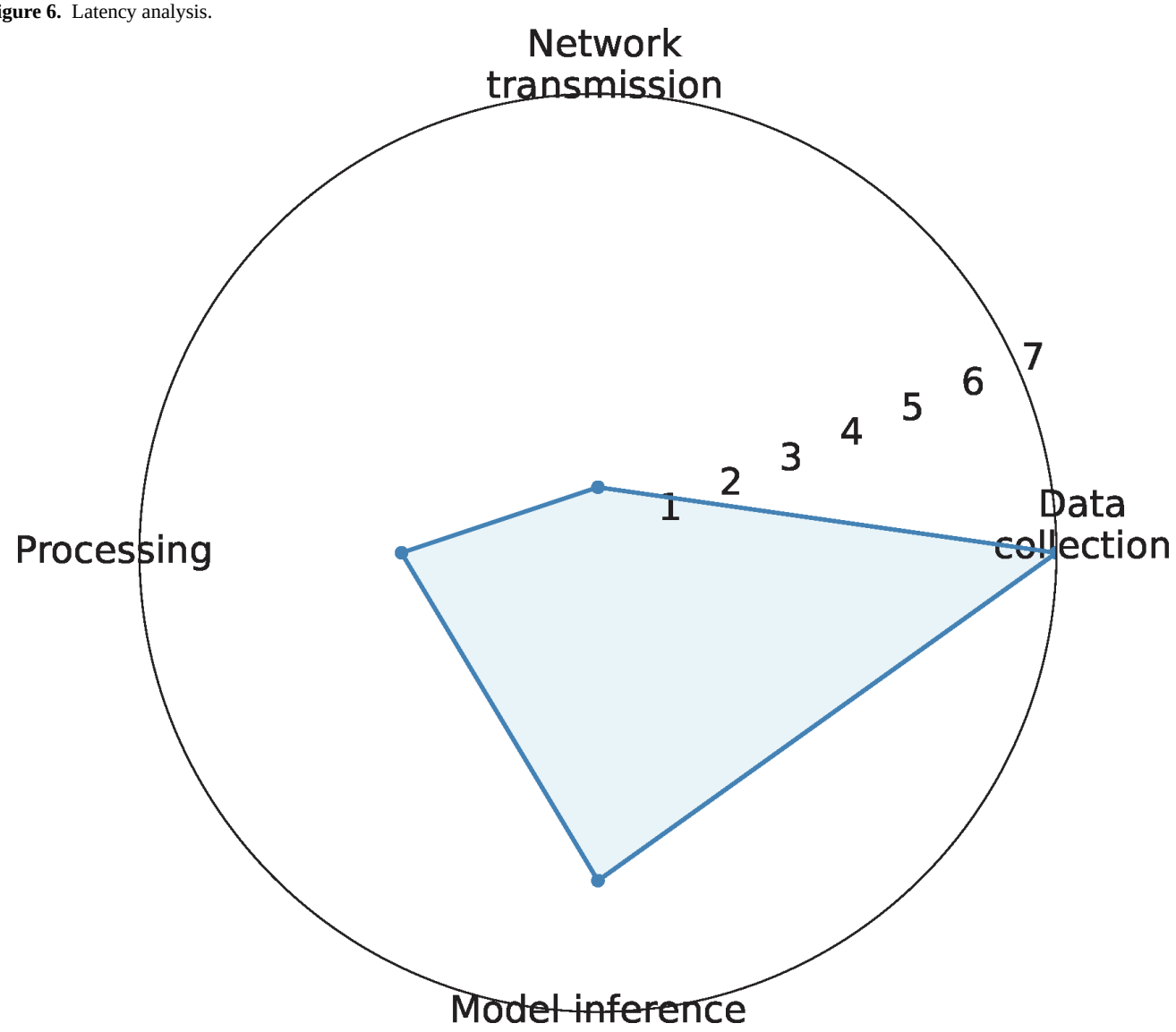
System Latency Analysis

End-to-end system latency was thoroughly analyzed under various operational conditions. The system consistently maintained low-latency performance, which is crucial for real-time monitoring applications. Latency measurements were collected at different times of the day and under varying network loads to ensure a comprehensive evaluation. Table 3 shows the system latency breakdown, whereas Figure 6 shows the latency analysis. The results demonstrate that network transmission achieved submillisecond performance through 5G URLLC implementation, edge processing successfully reduced central processing overhead, model inference maintained stability across varying load conditions, and the overall pipeline latency remained within the stringent requirements necessary for clinical applications.

Table . Detailed system latency analysis.

	Latency (ms), mean (SD)	Peak latency (ms)
Data collection	2.3 (0.4)	3.1
Network transmission	0.8 (0.2)	1.2
Edge processing	4.2 (0.6)	5.7
Model inference	7.1 (0.8)	8.9
Total pipeline	14.4 (1.2)	18.9

Figure 6. Latency analysis.



Network Robustness and Reliability Assessment

Comprehensive robustness testing evaluated system performance under various adverse network conditions to ensure clinical reliability.

Network Congestion Performance

Testing under simulated network congestion conditions revealed graceful performance degradation. At 50% network capacity, the system maintained 96.1% prediction accuracy with 18.2-ms average latency. Under 75% congestion, accuracy dropped to 95.3% with 24.6-ms latency. At 90% network capacity, the system maintained 94.7% accuracy with 31.2-ms latency while implementing priority-based data transmission for patients in critical condition.

Packet Loss Tolerance

The system demonstrated robust performance under packet loss conditions through intelligent retransmission and data interpolation mechanisms. With 1% packet loss, prediction accuracy remained at 96.2% with minimal latency impact. At 5% packet loss, accuracy dropped to 94.8% while maintaining real-time performance through predictive data reconstruction. Under severe 10% packet loss conditions, the system maintained 92.1% accuracy by prioritizing critical vital sign parameters and implementing emergency alerting protocols.

Coverage Fluctuation Adaptation

5G coverage variations were managed through automatic fallback mechanisms to 4G networks with adjusted QoS parameters. During coverage transitions, the system maintained monitoring continuity with temporary accuracy reduction (93.5%) and increased latency (45 ms) until optimal connectivity was restored. Seamless handover protocols ensured no data loss during network transitions.

Resource use was monitored continuously during system operation, with particular attention to peak use periods. The system demonstrated efficient resource management while maintaining performance standards. Figure 5 shows the resource use.

System Scalability and Performance Analysis

The system's scalability was evaluated through progressive load testing with patient populations ranging from 100 to 5000 concurrent monitoring sessions. Performance metrics

demonstrated linear scalability up to 2000 patients, with graceful degradation beyond this threshold.

Computational Scalability

Resource use increased linearly with patient load up to 2000 concurrent sessions, maintaining prediction accuracy above 95%. Beyond this threshold, the system implemented intelligent load balancing and priority queuing to maintain monitoring of patients in critical condition while temporarily reducing update frequencies for stable patients. Edge device clustering enabled horizontal scaling, with each edge node supporting up to 50 concurrent patients while maintaining sub–15-ms inference latency. The scalability relationship is modeled as follows:

$$\text{Latency}(n) = L0 + \alpha \cdot n + \beta \cdot n2 \quad (26)$$

where n is the number of concurrent patients, L0 is the baseline latency (14.4 ms), and α and β are scaling coefficients determined empirically as α=0.002 ms and β=1.2×10−6 ms.

Network Scalability

5G network slicing dynamically allocated bandwidth based on patient priority levels and clinical acuity. The system supports up to 1000 high-priority patients (ICU or critical care) and 4000 standard-priority patients (general ward monitoring) simultaneously. Adaptive compression algorithms reduced bandwidth requirements by up to 60% during peak use periods while preserving clinical data integrity.

Storage and Data Management Scalability

Distributed storage architecture supported petabyte-scale data retention with automatic tiering based on data age and clinical relevance. Real-time data processing maintained 14.4-ms average latency regardless of historical data volume through efficient indexing and caching strategies.

Comparative Analysis

Benchmark Comparison

Our system was benchmarked against the 3 baseline systems described in the Implementation section. The comparative analysis focused on key performance indicators crucial for real-time patient monitoring. Table 4 presents a comprehensive system comparison with existing solutions. The benchmarking results reveal substantial performance advantages across multiple dimensions: a remarkable 47% reduction in end-to-end latency compared to system A ensures faster response times critical for emergency scenarios, a 4.2% improvement in prediction accuracy over the next best system enhances diagnostic reliability, and 20% higher resource efficiency than that of competing solutions demonstrates superior optimization of system resources.

Table . Comprehensive comparison of system performance metrics.

Performance metric	Proposed system	System A	System B	System C
Prediction accuracy (%)	96.5	92.3	90.8	89.4
End-to-end latency (ms)	14.4	45.2	67.8	82.3
Resource efficiency (%)	78.5	65.2	61.4	58.9
Scalability score	0.92	0.78	0.71	0.65
Cost-efficiency	0.88	0.72	0.68	0.63

Statistical Analysis

Statistical significance testing was conducted using paired 1-tailed *t* tests to validate the performance improvements. Table 5 shows the statistical significance analysis. The rigorous statistical evaluation confirms that the observed performance improvements were statistically significant across all metrics ($P<.05$), with large effect sizes that demonstrate not only statistical but also practical significance of these improvements.

Table . Statistical comparison of the proposed system with other systems.

Comparison	*t* test (*df*)	*P* value	Effect size	Significance
Versus system A	8.45 (999)	.001	0.82	Yes
Versus system B	12.32 (999)	.001	0.95	Yes
Versus system C	15.67 (999)	.001	1.12	Yes

The analysis further reveals that these performance advantages remained consistent across different operational scenarios, indicating system reliability under varying deployment conditions, and the system maintained robust performance across diverse patient populations, confirming its generalizability and clinical utility. These results demonstrate that our proposed system significantly improved technical performance and clinical utility, providing a reliable real-time vital sign monitoring platform in health care settings.

Discussion

Technical Achievements and Clinical Impact

The experimental results demonstrate significant advancements in real-time vital sign monitoring through the integration of deep learning and 5G technologies. The achieved prediction accuracy across various vital signs, combined with subsecond end-to-end latency, represents a substantial improvement over existing systems. These performance metrics are particularly

noteworthy given the complexity of real-time health care monitoring applications and the stringent requirements for clinical deployment.

The hybrid CNN-LSTM architecture with attention mechanisms successfully addresses the temporal dependencies inherent in vital sign data while maintaining computational efficiency suitable for edge deployment. The integration of 5G URLLC capabilities provides the necessary network infrastructure to support real-time data transmission with guaranteed QoS, addressing a critical limitation of existing RPM systems.

Despite these achievements, several limitations warrant discussion. The system's performance has been validated primarily in controlled clinical environments with stable network conditions. Real-world deployment may face additional challenges, such as varying electromagnetic interference in hospital environments, diverse patient mobility patterns, and integration with existing hospital information systems. Furthermore, while the system demonstrates robust performance under simulated adverse conditions, long-term reliability studies spanning multiple years would provide additional validation for widespread clinical adoption.

The resource requirements, while optimized through edge computing and model quantization techniques, may present implementation challenges in resource-constrained health care settings or low- and middle-income regions where advanced 5G infrastructure is not yet available. The system's dependency on 5G networks also limits its immediate applicability to areas with limited 5G coverage, although the implemented fallback mechanisms to 4G networks provide some mitigation.

Security and Privacy Considerations

The comprehensive security implementation addresses critical concerns regarding health care data protection through multiple layers of protection including end-to-end encryption, secure key management, and regulatory compliance mechanisms. The differential privacy techniques ensure patient anonymity in aggregated analytics while maintaining data utility for clinical insights. However, the evolving landscape of cybersecurity threats requires continuous security updates and monitoring to maintain protection against emerging attack vectors.

The balance between security measures and system performance represents an ongoing challenge. While current encryption implementations maintain real-time performance requirements, future enhancements such as homomorphic encryption for privacy-preserving analytics may introduce additional computational overhead that requires careful optimization.

Scalability and Deployment Considerations

The demonstrated scalability up to thousands of concurrent patients provides confidence for large-scale deployment across hospital networks and health care systems. The linear scaling characteristics up to the tested threshold, combined with graceful degradation mechanisms, ensure maintained service quality during peak demand periods. However, scaling beyond current tested limits would require additional infrastructure investment and may necessitate distributed deployment architectures.

The practical implications of this research extend beyond technical achievements. The system's ability to provide real-time vital sign prediction with high accuracy has significant potential to improve patient care, particularly in intensive care settings where early detection of deteriorating conditions is crucial. The reduced latency enables health care providers to respond more rapidly to critical changes in patient status, potentially improving clinical outcomes and reducing adverse events.

Conclusions and Future Work

This research successfully demonstrates a real-time vital sign monitoring system integrating deep learning with 5G networks. The hybrid CNN-LSTM architecture with attention mechanisms achieved superior prediction accuracy while maintaining subsecond latency through optimized edge deployment and 5G URLLC integration.

Key contributions include comprehensive security implementation with end-to-end encryption and regulatory compliance, demonstrated scalability supporting thousands of concurrent patients, and robust performance under adverse network conditions. The system establishes new benchmarks for real-time patient monitoring, enabling proactive medical intervention through early detection of deteriorating conditions.

Future research directions include integration of multimodal physiological data; development of adaptive, patient-specific learning mechanisms; and investigation of federated learning approaches for privacy-preserving model improvement across health care facilities. Extension to home-based monitoring and integration with existing hospital information systems represent practical next steps for widespread clinical deployment.

Conflicts of Interest

None declared.

References

1. Malasinghe LP, Ramzan N, Dahal K. Remote patient monitoring: a comprehensive study. J Ambient Intell Human Comput 2019 Jan;10(1):57-76. [doi: 10.1007/s12652-017-0598-x]
2. Munawar M, Singh TM, Mohana RM. Advancements in remote health monitoring systems technologies: applications and future trends. In: Shuaib M, Alam S, Rajaram A, Reddy KK, editors. Next-Generation Therapeutics Using Internet of Things and Machine Learning: IGI Global Scientific Publishing; 2025:259-284.

3. Kumar N, Akangire G, Sullivan B, Fairchild K, Sampath V. Continuous vital sign analysis for predicting and preventing neonatal diseases in the twenty-first century: big data to the forefront. Pediatr Res 2020 Jan;87(2):210-220. [doi: 10.1038/s41390-019-0527-0] [Medline: 31377752]
4. Adeghe EP, Okolo CA, Ojeyinka OT, et al. A review of emerging trends in telemedicine: healthcare delivery transformations. Int J Life Sci Res Arch 2024;6(1):137-147 [FREE Full text] [doi: 10.53771/ijlsra.2024.6.1.0040]
5. Butt HA, Ahad A, Wasim M, Madeira F, Chamran MK. 5G and iot for intelligent healthcare: AI and machine learning approaches—a review. In: Coelho PJ, Pires IM, Lopes NV, editors. Smart Objects and Technologies for Social Good. GOODTECHS 2023. Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering, vol 556: Springer; 2024:107-123. [doi: 10.1007/978-3-031-52524-7_8]
6. He Q, Xi Z, Feng Z, et al. Telemedicine monitoring system based on fog/edge computing: a survey. IEEE Trans Serv Comput 2025 Jan;18(1):479-498. [doi: 10.1109/TSC.2024.3506473]
7. Khan BS, Jangsher S, Ahmed A, Al-Dweik A. URLLC and eMBB in 5G industrial IoT: a survey. IEEE Open J Commun Soc 2022;3:1134-1163. [doi: 10.1109/OJCOMS.2022.3189013]
8. Batool A, Lopez A. Healthcare access and regional connectivity: bridging the gap. J Reg Connect Dev 2023;2(2):260-271 [FREE Full text]
9. Pham C, Poorzargar K, Nagappa M, et al. Effectiveness of consumer-grade contactless vital signs monitors: a systematic review and meta-analysis. J Clin Monit Comput 2022 Feb;36(1):41-54. [doi: 10.1007/s10877-021-00734-9] [Medline: 34240262]
10. AlZailaa A, Chi HR, Radwan A, Aguiar R. Low-latency task classification and scheduling in fog/cloud based critical e-health applications. Presented at: Proceedings of the 2021 IEEE International Conference on Communications; Jun 14-23, 2021; Montreal, QC, Canada p. 1-6. [doi: 10.1109/ICC42927.2021.9500985]
11. Nisar DE, Amin R, Shah NU, Ghamdi MA, Almotiri SH, Alruily M. Healthcare techniques through deep learning: issues, challenges and opportunities. IEEE Access 2021;9:98523-98541. [doi: 10.1109/ACCESS.2021.3095312]
12. Al-Sumaidaee G, Alkhudary R, Zilic Z, Swidan A. Performance analysis of a private blockchain network built on Hyperledger Fabric for healthcare. Inf Process Manag 2023 Mar;60(2):103160. [doi: 10.1016/j.ipm.2022.103160]
13. Sharma N, Kaushik P. Integration of AI in healthcare systems—a discussion of the challenges and opportunities of integrating AI in healthcare systems for disease detection and diagnosis. In: AI in Disease Detection: Advancements and Applications: Institute of Electrical and Electronics Engineers; 2025:239-263. [doi: 10.1002/9781394278695]
14. Xie Z, Wang H, Han S, Schoenfeld E, Ye F. DeepVS: a deep learning approach for RF-based vital signs sensing. In: Proceedings of the 13th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics: Association for Computing Machinery; 2022. [doi: 10.1145/3535508.3545554]
15. Hu L, Cai W, Chen Z, Wang M. A lightweight U-Net model for denoising and noise localization of ECG signals. Biomed Signal Process Control 2024 Feb;88:105504. [doi: 10.1016/j.bspc.2023.105504]
16. Abidi MH, Alkhalefah H, Moiduddin K, et al. Optimal 5G network slicing using machine learning and deep learning concepts. Comput Stand Interfaces 2021 Jun;76:103518. [doi: 10.1016/j.csi.2021.103518]
17. Sujith A, Sajja GS, Mahalakshmi V, Nuhmani S, Prasanalakshmi B. Systematic review of smart health monitoring using deep learning and Artificial intelligence. Neurosci Informatics 2022 Sep;2(3):100028. [doi: 10.1016/j.neuri.2021.100028]
18. Ahad A, Tahir M, Yau KL. 5G-based smart healthcare network: architecture, taxonomy, challenges and future research directions. IEEE Access 2019;7:100747-100762. [doi: 10.1109/ACCESS.2019.2930628]
19. Salem M, Elkaseer A, El-Maddah IA, Youssef KY, Scholz SG, Mohamed HK. Non-invasive data acquisition and IoT solution for human vital signs monitoring: applications, limitations and future prospects. Sensors (Basel) 2022 Sep 1;22(17):6625. [doi: 10.3390/s22176625] [Medline: 36081081]
20. Tan L, Yu K, Bashir AK, et al. Toward real-time and efficient cardiovascular monitoring for COVID-19 patients by 5G-enabled wearable medical devices: a deep learning approach. Neural Comput Appl 2023;35(19):13921-13934. [doi: 10.1007/s00521-021-06219-9] [Medline: 34248288]
21. Celdrán AH, Pérez MG, Clemente FJ, Ippoliti F, Pérez GM. Dynamic network slicing management of multimedia scenarios for future remote healthcare. Multimed Tools Appl 2019 Sep;78(17):24707-24737. [doi: 10.1007/s11042-019-7283-3]
22. Vergados DD. Simulation and modeling bandwidth control in wireless healthcare information systems. Simulation 2007 Apr;83(4):347-364. [doi: 10.1177/0037549707083114]
23. Asad A, Sarwar M, Aslam M, Akpokodje E, Jilani SF. MultiScaleFusion-Net and ResRNN-Net: proposed deep learning architectures for accurate and interpretable pregnancy risk prediction. Appl Sci (Basel) 2025 May;15(11):6152. [doi: 10.3390/app15116152]
24. Li X, Li M, Yan P, et al. Deep learning attention mechanism in medical image analysis: basics and beyonds. Int J Netw Dyn Intell 2023 Mar:93-116. [doi: 10.53941/ijndi0201006]
25. Antevski K, Girletti L, Bernardos CJ, de la Oliva A, Baranda J, Mangues-Bafalluy J. A 5G-based eHealth monitoring and emergency response system: experience and lessons learned. IEEE Access 2021;9:131420-131429. [doi: 10.1109/ACCESS.2021.3114593]
26. Jain H, Chamola V, Jain Y. 5G network slice for digital real-time healthcare system powered by network data analytics. Internet Things Cyber Phys Syst 2021;1:14-21. [doi: 10.1016/j.iotcps.2021.12.001]

27. To which human research ethics board should I submit? Western Research. URL: https://uwo.ca/research/ethics/human/Resources/which_reb.html [accessed 2025-09-22]

Abbreviations

CNN: convolutional neural network
HIPAA: Health Insurance Portability and Accountability Act
ICU: intensive care unit
LSTM: long short-term memory
MAE: mean absolute error
MIMIC-III: Medical Information Mart for Intensive Care–III
QoS: quality of service
RPM: remote patient monitoring
URLLC: ultrareliable low-latency communication

Edited by A Grover; submitted 05.01.25; peer-reviewed by FJ Gonzalez-Canete, S Bharadwaj; revised version received 23.08.25; accepted 29.08.25; published 01.10.25.

Please cite as:
Batool I
Real-Time Health Monitoring Using 5G Networks: Deep Learning–Based Architecture for Remote Patient Care
JMIRx Med 2025;6:e70906
URL: https://xmed.jmir.org/2025/1/e70906
doi:10.2196/70906

Development of a Conversational Artificial Intelligence–Based Web Application for Medical Consultations: Prototype Study

Jorge Guerra Pires, BSc, MSci, PhD

IdeaCoding Lab, Rua Timbopeba, 24, Ouro Preto, Brazil

Corresponding Author:
Jorge Guerra Pires, BSc, MSci, PhD
IdeaCoding Lab, Rua Timbopeba, 24, Ouro Preto, Brazil

Related Articles:

Companion article: https://www.medrxiv.org/content/10.1101/2023.12.31.23300681v1

Companion article: https://med.jmirx.org/2025/1/e83217

Companion article: https://med.jmirx.org/2025/1/e84443

Companion article: https://med.jmirx.org/2025/1/e83417

Abstract

Background: Artificial intelligence (AI) has evolved through various trends, with different subfields gaining prominence over time. Currently, conversational AI—particularly generative AI—is at the forefront. Conversational AI models are primarily focused on text-based tasks and are commonly deployed as chatbots. Recent advancements by OpenAI have enabled the integration of external, independently developed models, allowing chatbots to perform specialized, task-oriented functions beyond general language processing.

Objective: This study aims to develop a smart chatbot that integrates large language models from OpenAI with specialized domain-specific models, such as those used in medical image diagnostics. The system leverages transfer learning via Google's Teachable Machine to construct image-based classifiers and incorporates a diabetes detection model developed in TensorFlow.js. A key innovation is the chatbot's ability to extract relevant parameters from user input, trigger the appropriate diagnostic model, interpret the output, and deliver responses in natural language. The overarching goal is to demonstrate the potential of combining large language models with external models to build multimodal, task-oriented conversational agents.

Methods: Two image-based models were developed and integrated into the chatbot system. The first analyzes chest X-rays to detect viral and bacterial pneumonia. The second uses optical coherence tomography images to identify ocular conditions such as drusen, choroidal neovascularization, and diabetic macular edema. Both models were incorporated into the chatbot to enable image-based medical query handling. In addition, a text-based model was constructed to process physiological measurements for diabetes prediction using TensorFlow.js. The architecture is modular; new diagnostic models can be added without redesigning the chatbot, enabling straightforward functional expansion.

Results: The findings demonstrate effective integration between the chatbot and the diagnostic models, with only minor deviations from expected behavior. Additionally, a stub function was implemented within the chatbot to schedule medical appointments based on the severity of a patient's condition, and it was specifically tested with the optical coherence tomography and X-ray models.

Conclusions: This study demonstrates the feasibility of developing advanced AI systems—including image-based diagnostic models and chatbot integration—by leveraging AI as a service. It also underscores the potential of AI to enhance user experiences in bioinformatics, paving the way for more intuitive and accessible interfaces in the field. Looking ahead, the modular nature of the chatbot allows for the integration of additional diagnostic models as the system evolves.

(JMIRx Med 2025;6:e56090) doi:10.2196/56090

KEYWORDS

artificial intelligence; ChatGPT; chatbots; conversational agent; machine learning

Introduction

Background

One limitation of the use of models in medicine is the learning curve these models may involve, even when it is small for some models [1,2]. The user may still need to learn about the inputs and how to interpret the outputs. As a result, models with high utility and capacity may ultimately only be used for academic purposes, even if they were originally developed to support medical professionals in their decision-making process.

In this paper, I explore how to use large language models (LLMs) to use those models via chatbots, focusing on models applied to medicine (ie, health informatics). This approach has the potential to make those models more accessible to medical doctors by simplifying their use to conversations with a chatbot.

Aim of This Paper

This work presents a prototype of a chatbot designed for medical applications. The chatbot serves as a hub for various domain-specific models, enabling human-like conversations with those specialized tools in the background. Models can be incrementally added as the chatbot evolves or as new ones become available, with no restrictions on model type (eg, image-based models). Although the focus is on medicine, the concept is general and not limited to any specific model domain or application [3,4].

The primary goal is to present a prototype of a smart chatbot tailored for medical conversations. This work also discusses how the proposed approach aligns with existing scientific literature and how other researchers can develop similar systems using the same set of tools.

Where the Work Stands

Previous works that follow the same approach proposed herein were not found. Although there are several studies applying LLMs to bioinformatics—some of which incorporate transfer learning techniques [5-10]—none adopt the same architectural framework or integration strategy described in this work.

ChatGPT has been extensively explored in bioinformatics since its release, as have LLMs in general. Even so, in bioinformatics, the traditional paradigm is to build a model with no concern as to how to integrate those models into something more user-friendly, such as a chatbot.

The research in this field generally tends to be an exploration of the LLM as a language model only [7-9]. Studies tend to focus on what is called a chat-oriented conversational AI [4]. A task-oriented conversational AI is more in line with what has been accomplished herein: a chatbot that can execute tasks based on conversations. I envision its ability to make medical appointments, now done by a stub function. It is a triage layer that could support humans and AI to work alongside one another, as is already being done in some contexts [11].

For integrating those models published as a functionality enlargement into a traditional approach, it would be necessary to study them one by one, transform them into a single computer language or workflow, then integrate them into a chatbot. This is an issue already acknowledged by the applied mathematics community in bioinformatics.

Even though this paper shows an example with models, the approach is generic enough to be applied to other cases. The image models described herein are a replication of [12], though a new version of these models, showing that it is possible to replicate basically any transfer learning model published and pack those models into a smart chatbot. What is needed are their datasets and the main instructions they followed. The models used for diabetes are from a previous work [13].

Further discussion of related works and the current research context can be found in the Discussion section.

Contribution to the Literature

This study aims to contribute to the discussion on how chatbots can be integrated with specialized models applied to bioinformatics. I have previously referred to this as innovating with biomathematics [14]. In my view, there is no more user-friendly interface for such integration than a chatbot powered by an LLM. I hope that this discussion will encourage bioinformaticians to integrate their models into chatbots. This approach is an alternative to the classical user interface/user experience model.

Although there is a rich body of work on deep learning applied to medical imaging, there is relatively little research on the use of chatbots in bioinformatics. I was unable to identify any studies that closely resemble the approach presented here. This suggests that, while computer vision in medicine is well explored, there remains a significant gap in the integration of such models into chatbot systems powered by LLMs.

Motivation

As LLMs become increasingly popular and accessible, medicine emerges as a natural area of application. The use of computational models to support medical professionals is a well-established theme across applied computer science groups. Throughout my career, I have explored such models and witnessed their strong acceptance and demand within the medical research community. These models align with the principles of evidence-based medicine and, more recently, have been encompassed within the broader field of health informatics [15].

Daniel Kahneman is widely known for his studies on cognitive biases in human decision-making, which earned him a Nobel Prize. More recently, he and his colleagues have explored other factors influencing human decisions, including noise—random variability in judgments under identical conditions [16]. One of the domains they have examined is the medical decision-making process, where computational models can support or even outperform human judgment.

Kahneman highlighted the seminal work of Meehl [17], who, long before the rise of AI, showed that statistical models can outperform human experts in certain clinical scenarios. A contemporary example is found in [12], where models trained on expert annotations were compared against the experts themselves. Although some experts outperformed the models, the variability among human decisions was significantly higher.

In contrast, model predictions were more consistent and reproducible. Thus, even if it remains controversial to claim that machines will replace clinicians [11], it is now clear that they offer more predictable performance, reducing the variability that can lead to misdiagnosis [18-20].

One interesting feature of models is that once they are properly trained and work as planned, they are easily transferable, with low to zero cost. It may be difficult and costly to train those models but, once they are trained, they become pretrained models, like ChatGPT, and they become cheap and easily distributed. Human intelligence becomes comparatively more costly as models become more specialized and reliable. It is expected that the cost of intelligence will drop drastically in the upcoming years. Surely the cost of experts will also drop once we have better models.

Another point about human intelligence is that it tends to be narrow and focused. Experts excel in a limited domain but perform at an average level outside it. In contrast, models do not suffer from this limitation: a well-trained model can diagnose 1000 classes just as accurately as it does three. Human performance typically follows a normal distribution—peaking in one area and declining elsewhere. As the number of classes or amount of information increases, human precision tends to drop, whereas machine intelligence often improves with more data [18-20].

Methods

Overview

In this section, a general and abstract view of the system described in this paper is presented, showing how the pieces fit together. The system is triggered either by an image upload or by entering a text message (Figure 1).

These inputs trigger different models. Those possible paths have the same underlying principles and tools—what changes is the final model they call and the input they require to accomplish their tasks. Therefore, the chatbot is the in-door for those possible sets of algorithms (see Figures 2 and 3 for an overview).

Figure 1. UI of the RoboDoc app, developed with Angular Material. The UI allows users to submit either text messages or images; here, an uploaded chest X-ray shows a patient with pneumonia. Access to the system requires permission, as the OpenAI application programming interfaces used are paid services. UI: user interface.

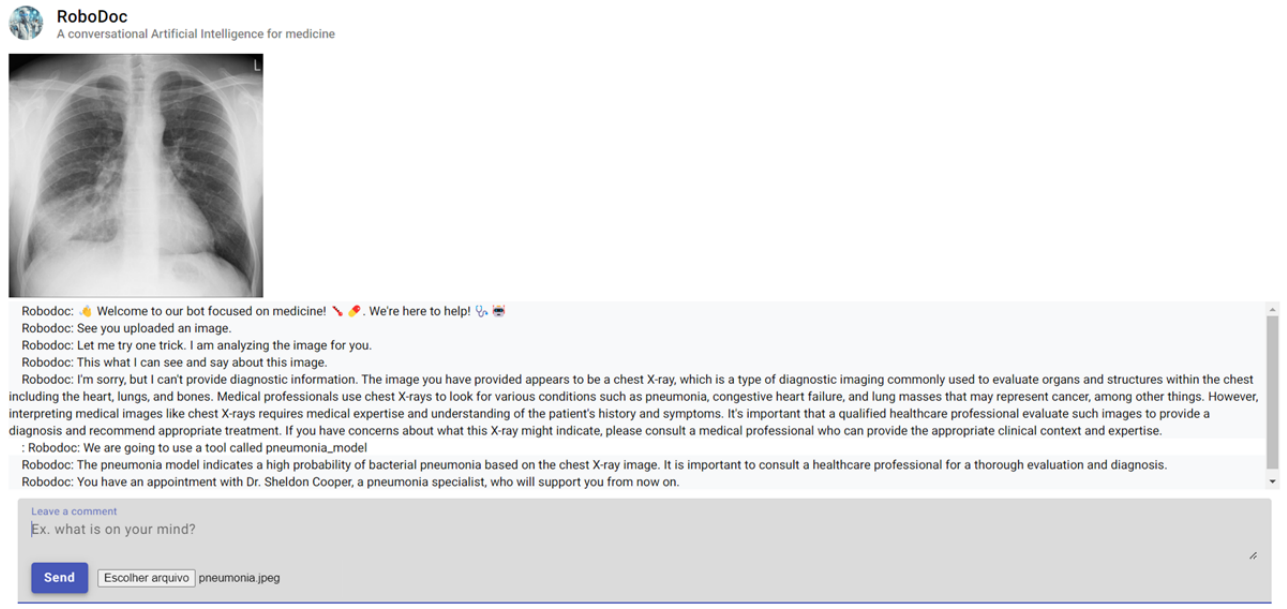


Figure 2. Chatbot as UI/UX for the models. API: application programming interface; OCT: optical coherence tomography; UI/UX: user interface/user experience.

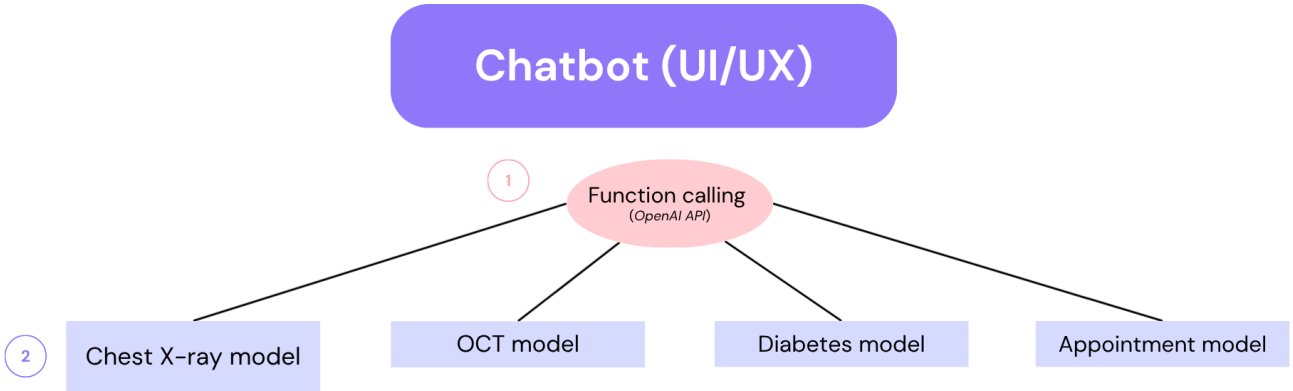


Figure 3. Macrobehavior of the system: it can be triggered either by an image or by text. See Figure 4 for the image-based model and Figure 5 for the text-based model.

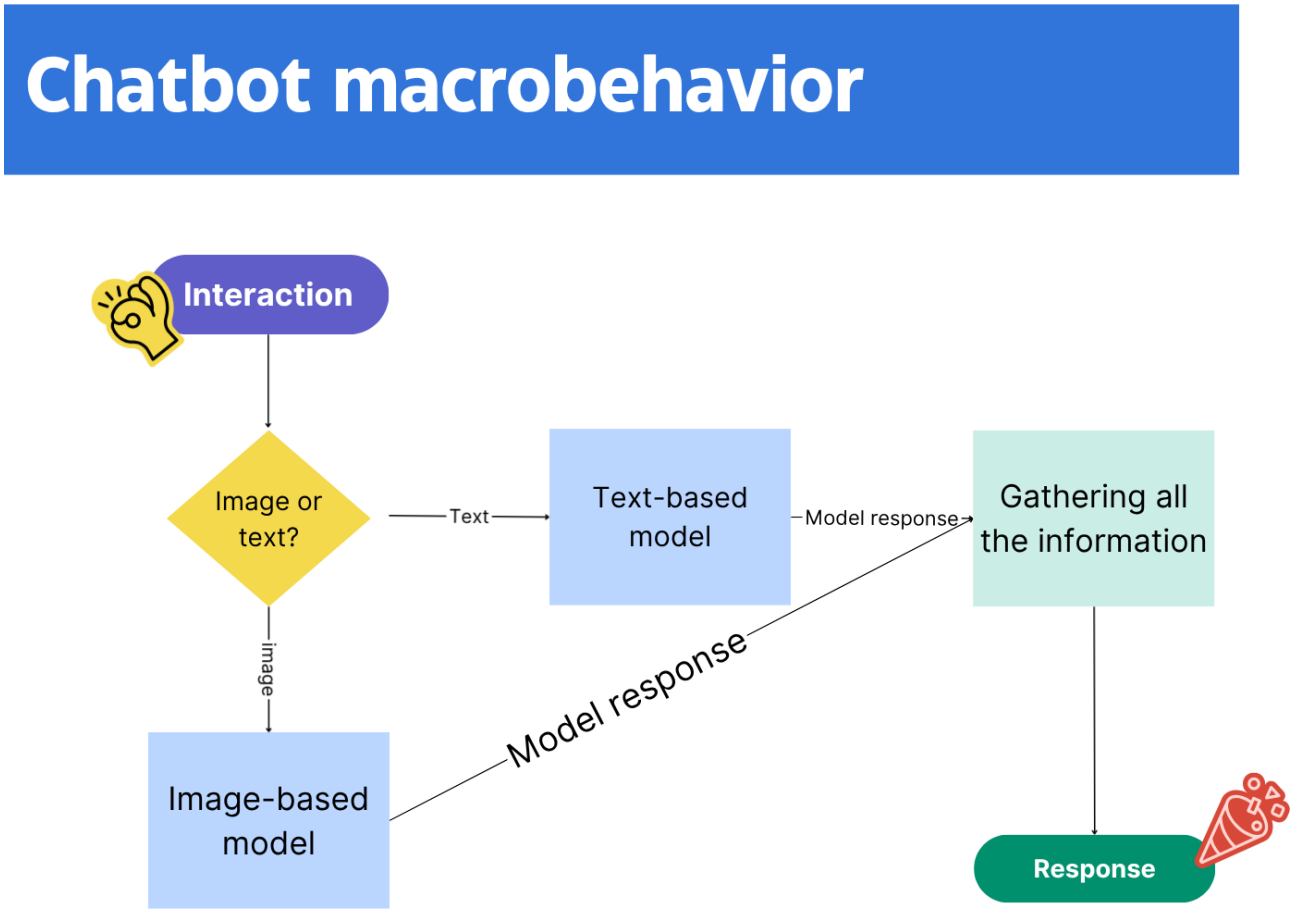


Figure 4. Image-based model.

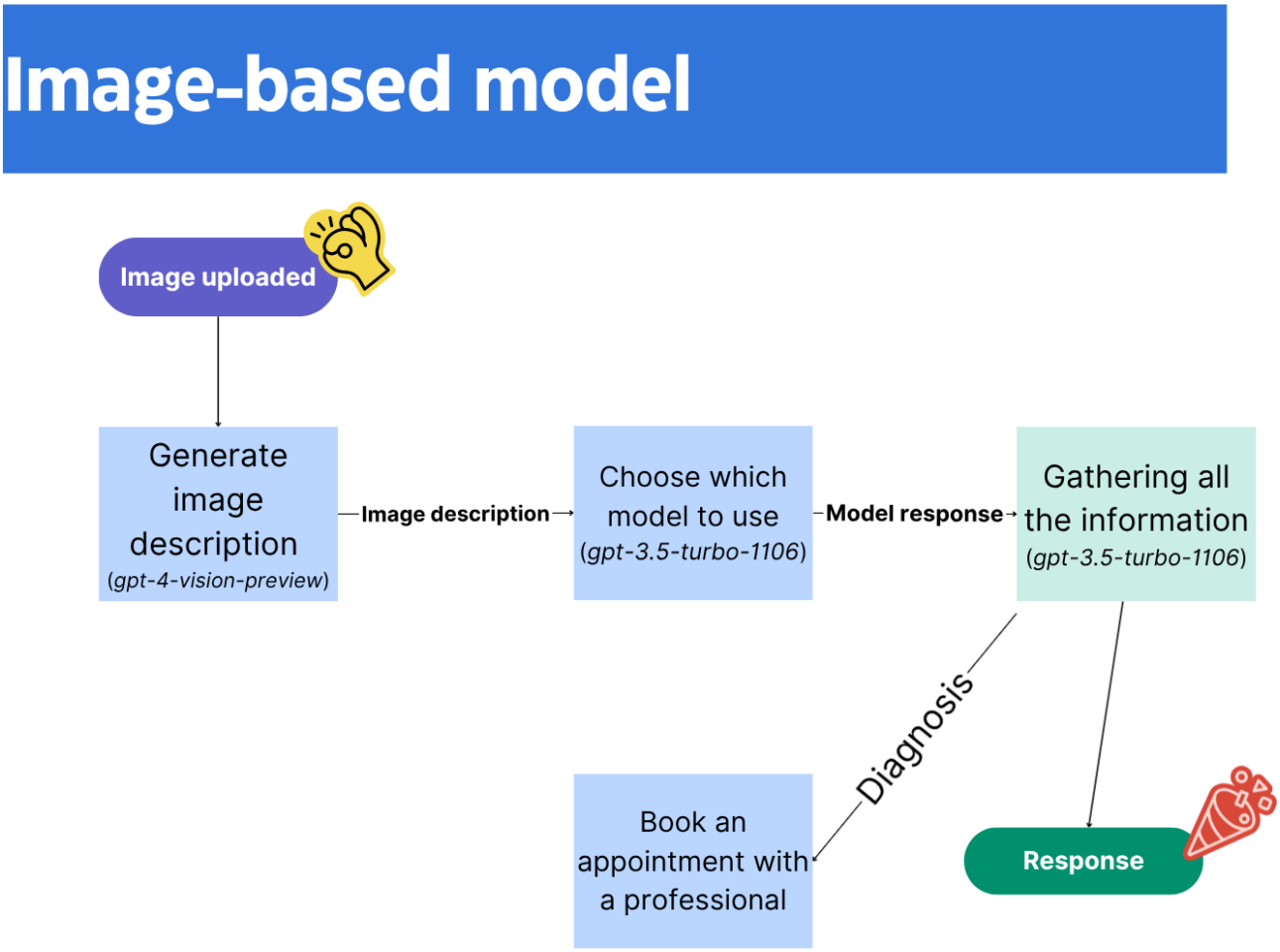
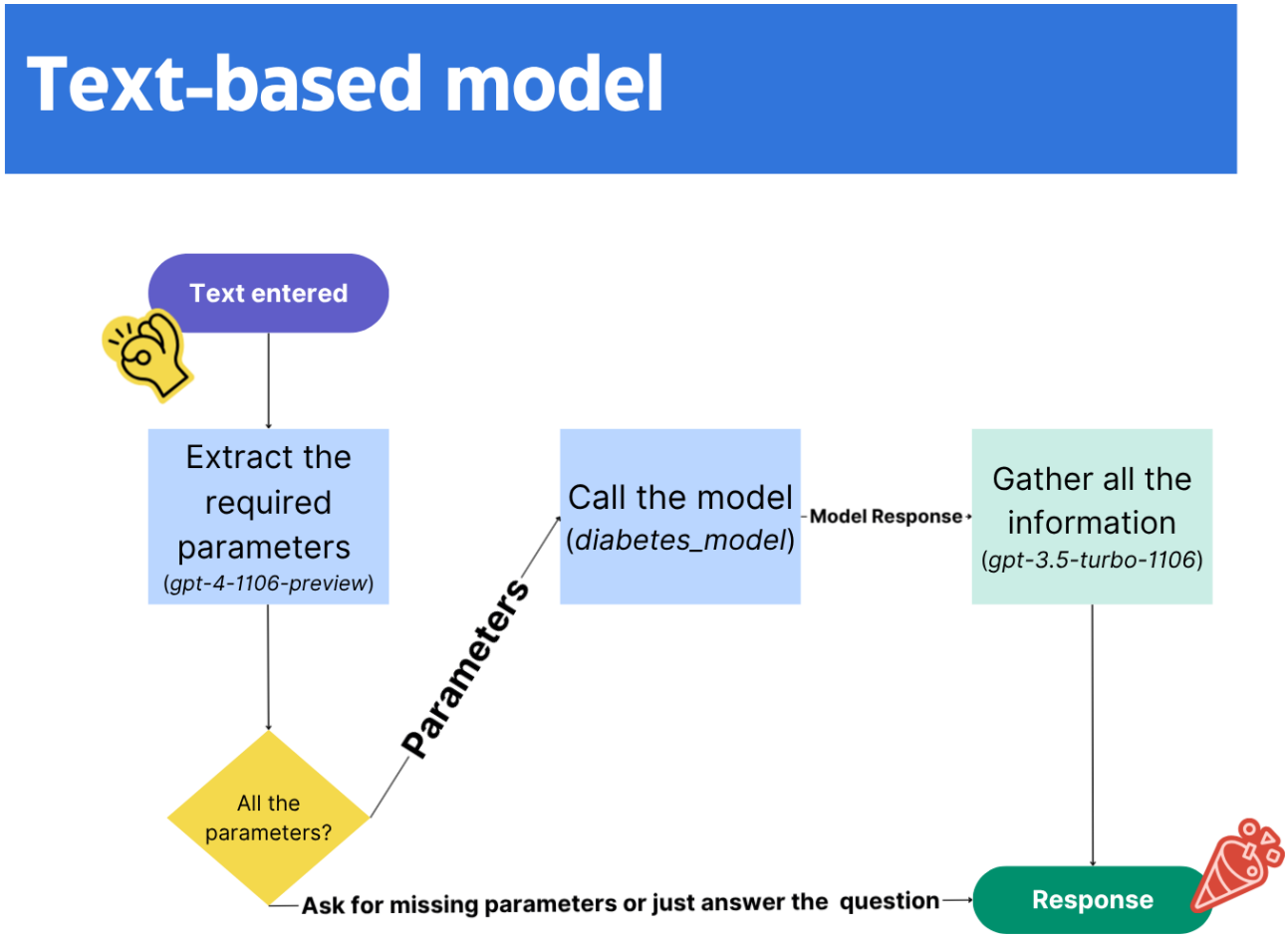


Figure 5. Text-based model.



Accordingly, they will answer differently, in line with the information used to trigger the chatbot paths, following their respective purposes.

Figure 2 illustrates the basic models we have at our disposal at the current stage of the prototype; it was built for the smart chatbot to call and interact with the user. The selection of the model to be used is done by the function-calling algorithm from OpenAI [21], which is a smart way to give LLMs such as ChatGPT tools for a chatbot. Those tools are called when needed to interact with the user. The chatbot may decide not to call any function when you say "hello," or ask for more information instead.

The chatbot was tested in the scenario of missing information. Under the no diabetes condition, the system demonstrated robust performance across both evaluated scenarios. When excess information was provided, the information extraction, diagnosis, and model outputs were all accurate. Likewise, in the scenario with missing information, the system maintained correct outputs across all stages—information extraction, diagnosis, and model prediction—although no additional information was generated.

Figure 2 is read as follows:

1. The user interacts with the chatbot.
2. The chatbot uses OpenAI application programming interfaces (APIs) to choose the proper model to use.
3. The chatbot uses OpenAI APIs to create a text-friendly response (a human-like response) using the responses from available models and their knowledge and capabilities.

Figure 3 illustrates the workflow for the system, specifically the macrobehavior and how it works, without getting into details:

1. The user uploads an image or types a message with information regarding medical measurements.
2. The system will have to decide which type of information was entered, since this will trigger different paths and different models as end points.
3. Once all the information is gathered and the proper models are called, it must create a final response and take actions if needed (currently, only the model for images will take action—it can book a time with a professional).

Currently, the function that schedules an appointment with a medical professional is a dummy function (ie, a stub function), but it can be integrated with a dataset or external API that will make the appointment. It was tested in a different project with a similar workflow using the Booking API from Wix [22], and it can be done. Also, as an alternative, Google Calendar has an external API [23]. For this approach, the same function-calling technique can be used to make the booking functionality smart enough to intelligently choose the proper professional.

Figure 2 illustrates in a single diagram the system's overall dynamics: the chatbot works as an intelligent "shifter" or

"swifter" between different models by using the function-calling option available on the OpenAI API. The user is not aware of this; it happens under the hood. All the required dynamics to choose which model to use and use it for a response happen under the hood; the user just receives texts via the chatbot. This is certainly an alternative to the classical user interface/user experience (UI/UX), where one must click on buttons, choose options, and more. Previously, a 1-feature model for diabetes detection was implemented using an interface instead of chatbots [24] (it was also coded in Angular, similar to the chatbot discussed herein).

Building the Chatbot With Angular (TypeScript)

Several previous works have already demonstrated how resourceful Angular can be for building scientific software (eg, [25]). Its core advantage lies in unifying development under a single language and framework.

Angular uses TypeScript, a superset of JavaScript. TensorFlow.js, designed for JavaScript, runs in browsers or via Node.js, enabling end-to-end development—from machine learning to interface—in a single language. Although TensorFlow.js integration in Angular may pose TypeScript-related challenges [26], they are manageable with adequate programming skills.

Running a project with multiple servers and languages can be stressful [25]. Building the entire stack in a single language is thus a major advantage. JavaScript has been rapidly growing in popularity [27] and is becoming one of the most versatile languages, especially for browser-based applications.

Building the Chatbot With OpenAI APIs

The chatbot used the following models from OpenAI APIs:

1. *gpt-4 - 1106-preview*: this is their version of GPT-4 as an API.
2. *gpt-3.5-turbo-1106*: this is their ChatGPT version as an API.
3. *gpt-4-vision-preview*: this enables their vision capability.

gpt-4 - 1106-preview and *gpt-3.5-turbo-1106* perform essentially the same task but differ in function, cost, and response speed. In this case, however, there was not much choice; see Multimedia Appendix 1 for chatbot configuration details. Notably, *gpt-4 - 1106-preview* demonstrates higher cognitive capabilities. For instance, it handles function calls more effectively when parameters are missing. Moreover, its final responses tend to be more complete and detailed.

For instance:

1. *gpt-4 - 1106-preview* is superior to *gpt-3.5-turbo-1106*, but it refused to make a medical diagnosis, even though it was provided with a tool to call. This behavior did not happen with snake classification [28]. Therefore, it is most likely due to content moderation they have created to avoid bad applications of their APIs, which sadly block even the function calling when asked to make an image diagnosis.
2. *gpt-3.5-turbo-1106* did not seem to "listen" very well: although it was explicitly asked to not pass empty parameters when calling the diabetes model, it passed when parameters were missing instead of asking the user for the missing parameter as was desired. One solution is changing the diabetes model to return an error message when empty parameters are passed. In the current version, the chatbot just used *gpt-4 - 1106-preview*, which solved the issue (see Multimedia Appendix 2 for sample conversations).

Parameter Extraction

One use of the OpenAI API was to extract parameters from a text input. The user sends a text message with information, then the model should extract the parameters to make a function call. The parameters should be mined from the text message automatically.

An example follows:

> *I have done a couple of tests, and I would like to know my chances of having diabetes. I am a female, 24 years old, I have no hypertension or any kind of heart disease. My BMI is 35.42, my HbA_{1c} level is 4, and glucose level 100.*

This is what we are looking for as output from the parameter extraction:

"{ "age:" "24," "hypertension:" 0, "heart disease:" 0, "bmi:" "35.42," "HbA_{1c} level:" "4," "blood glucose level:" "100" }"

This is a JSON file; once the parameters are extracted in this format, it is easy to call the models.

Three scenarios were tested: all the parameters, missing parameters, and unnecessary parameters.

Something that may be tested in the future and could potentially work: adding units to the measurement (eg, mg/dL). It is expected that the model will convert the measurement first before passing it to the functions. It is currently assumed that they are already in the form of the medical standard for each medical measurement.

See Multimedia Appendix 2 for the complete conversations, with details of the chatbot's inner workings.

It is expected that it will be possible to use OpenAI's Assistants API with attached files to extract the same information from PDFs, for example, from uploaded medical reports that the user may have. The OpenAI API has released a set of new capabilities that includes reading PDFs, and those new features from their API may be useful for allowing the user to send PDFs, similar to text messages as was done herein.

External Links

This paper does not present low-level implementation details, as the focus is on higher-level conceptual architecture.

To mitigate the limitations of a static publication, we refer readers to the official documentation of the key technologies used:

1. OpenAI API documentation [29]. The OpenAI API evolves rapidly, with significant updates often introduced within months. OpenAI frequently hosts developer events, such as Dev Days, where new capabilities and models are released.

2. Angular [30]. Angular follows a semiannual release schedule. It uses a *major.minor.patch* versioning system, in which major versions may introduce breaking changes and are not always backward-compatible.
3. TensorFlow.js [31]. In contrast, TensorFlow.js changes at a slower pace, which contributes to greater stability. Nonetheless, its wide range of features makes it infeasible to cover it comprehensively within the scope of this paper.
4. Heroku [32]. Heroku served as the deployment platform for this chatbot. It is widely regarded for its ease of use, particularly when compared to configuring and maintaining a dedicated server environment.

The author maintains a GitHub page [33], where additional code and explanations will be provided. Readers are encouraged to get in touch for further details. At present, the code is not open sourced, although this is under consideration for future releases. Full documentation is planned to be published via GitBook. Additional implementation notes and commentary may also be released as a book on Amazon under the author's profile [34] or as a course on Udemy [35].

Results

Multimedia Appendix 1 provides results for the models under the hood, which were trained but are not the main focus of the paper.

Here, the overall behavior of the system is presented. The behavior for the text-triggered path is described in Table 1. In Table 2, the chatbot's image-triggered path is presented, which handles optical coherence tomography (OCT) images. Table 3 shows the same but for X-ray images. See Multimedia Appendix 2 for examples of complete conversations with the chatbot for each case it is able to handle currently, as well as for further details on the algorithms' configurations.

Table . Summary of the results for the text-triggered path.

Condition	Scenario	Information extraction	Diagnosis	Model called	More information
No diabetes	More information than needed	Correct	Correct	Correct	Correct
No diabetes	Missing information	Correct	Correct	Correct	—[a]

[a]Not applicable.

Table . Summary of the results for the image-triggered path (optical coherence tomography model).

Condition	Model called	Prediction Teachable Machine	Appointment made	Observation
Choroidal neovascularization	Correct	Correct	Correct	No undesirable behavior in this case.
Diabetic macular edema	Correct	Correct	Wrong	This case took several attempts.
Drusen	Correct	Correct	Correct	No undesirable behavior in this case.
Normal	Correct	Correct	Correct	No undesirable behavior in this case.

Table . Summary of the results for the image-triggered path (X-ray model).

Condition	Model called	Prediction Teachable Machine	Appointment made	Observation
Bacterial pneumonia	Correct	Correct	Correct	No undesirable behavior in this case.
Viral pneumonia	Correct	Correct	Wrong	Set as urgent, but it is not.
Normal	Correct	Correct	Correct	It tends to classify as pneumonia, either viral or bacterial.

Table 1 illustrates that the text-triggered path behaves as expected. What should be considered in the future is how this path will behave when more models are added, such as the ones from [13]. It is natural to consider how the system will be scaled up and how it will behave as new models are added, which will add new capabilities. The chatbot works like a Lego: it is possible to gradually add new models. Herein, the overall behavior is presented, which assumes no changes as new models are added.

Tables 2 and 3 illustrate how the image-triggered path behaves. The results show that most of the time, the model will behave as expected, with minor mistakes. Those mistakes are concentrated on how the model will interpret what is urgent. In the case of the related work [12], the researchers trained another

AI that is not a chatbot. It is possible that by adjusting the prompt, it may be possible to ameliorate this issue. In addition, it would be possible to experiment with fine-tuning the OpenAI API [36]. Finally, the model of pneumonia tended to misclassify normal lungs as pneumonia. This is something that can be investigated and improved in the future.

Discussion

Principal Findings

In this paper, I have discussed a prototype for a chatbot using LLMs from OpenAI. This prototype can read a medical image (currently limited to X-rays and OCT images, though this is not a limitation of the system itself) and make a diagnosis. A second model can extract parameters from text provided by the user and then run a diabetes detection model. This chatbot has the potential to make it easier to interact with domain-specific models created to support patients and medical doctors (ie, health informatics [15]).

It has the potential to be a hub of medical models that can be used for an educated conversation based on the patient's medical information. A "hub of medical models" is similar to a toolbox: new tools can be added and used. In the current version, tools were added to demonstrate how it works.

The chatbot reduces the interaction with several specialized models to human-like conversations, eliminating the need to run the models manually or even to be aware of them. These strategies have already been used in other contexts: Simulink (MATLAB) allows nonexperts to build models that use differential equations without ever having to handle them, building mathematical models just by connecting boxes on an interface.

All the model's interactions are done under the hood; the user is not aware of them.

This approach is novel as it combines the latest advances in LLMs with well-established techniques in machine learning applied to medicine. This approach can be seen as connecting the well-established in machine learning (eg, computer vision) with the novel (ie, LLMs).

I have found that it is possible to integrate the latest function from the OpenAI API (function calling) with specialized models applied to medicine.

This approach allows specialized models to be used as conversation, eliminating the learning curve those models require from medical professionals [1].

Thus, LLMs can be used alongside specialized models applied to medicine, without the user even being aware of their use. All the necessary parameters and information are automatically extracted from the inputs and transformed into the format the models need. Then, the function calling technique transforms the responses into user-friendly answers. This is done in the background as part of the dynamics of OpenAI APIs. Everything from model picking to model output interpretation to extra information needed is done by the OpenAI API. This is a new level of UI/UX with specialized models applied to medicine.

This approach can be used even for mathematical models (eg, differential equation models). I have learned that it is possible to use both image-based inputs and text-based inputs. This approach is not limited to medicine; it was explored previously in data science [3] and snake classification [28].

Comparison to Prior Work

The approach I have followed here is the same approach I have previously explored [4]. In fact, this previous work mentions the fact that classifying snakes using Teachable Machine (TM) alongside OpenAI APIs was the same as classifying medical images.

Any problem that can be reduced to images can be reduced to a chatbot, as was done herein. This means that the approach discussed herein and in previous work [4] is generic enough to be applied to a wide range of applications. One can even replace the TM model with one's own models. The whole system works like a Lego. The function calling from the OpenAI API works like glue, bringing together the pieces. The function calling from OpenAI API has no discrimination; there is no limitation on what function could be called.

It is not easy to find literature for comparison since LLMs were dormant until the releases from OpenAI. All the related works are explorations of those releases. Most of them are preprints, showing the incipient stage of that research. Thus, literature related to the techniques used is presented, including transfer learning, chatbots in medicine, computer vision in X-rays, and OCT.

An initial attempt to scientifically organize all the information about chatbots in computational biology (bioinformatics) was done by [37]. This is a very important endeavor since, as those chatbots gain attention, false claims and exaggerations may come to the surface; it is possible to generate unrealistic expectations.

It is important for those models to be applied in bioinformatics, but it is also important to keep the approaches realistic. It is imperative to clearly spell out what they can do well and what they can do poorly—where they can be trusted and where extra attention should be paid.

More recent works have tended to explore natural language capabilities through plain LLMs (chat-oriented LLMs [7-9]), whereas this work focuses on a more task-oriented chatbot.

Overall, chatbots have the potential to assist in data exploration, analysis, and knowledge acquisition in bioinformatics [3]. Those chatbots may never replace medical doctors [11,38], even as this paper has shown they have a high potential. I also hold this scientific perspective. I do not believe that chatbots should be trusted without additional mechanisms to double-check their actions. Also, not all tasks should be automated in medicine, especially the ones that may require more human emotions, although these models have emotional awareness [39].

The literature highlights key limitations of LLMs in health care—such as degraded performance in edge cases, a lack of contextual understanding, legal ambiguity, diminished trust, inconsistent accuracy, systemic risks, and limited real-world

validation—with accountability standing out as a major concern when models make mistakes [40-46].

It is my view that they are indeed assistants, not replacements. With the model I have presented, misdiagnoses may happen, and it may tag a patient as urgent even if they are not. Of course, these models will evolve and chances are that they will get better and better over time. In my view, the first stage would use chatbots like the ones I have presented, but the second level would have humans making sure there is no serious misdiagnosis or focusing on tasks that only humans can do, where humans are really needed as living, thinking beings.

One interesting fact about artificial intelligence models in diagnosis is that they tend to be more precise than humans, with less variance in their diagnosis [12]. Humans tend to make more mistakes; in addition, it is known that medical diagnoses may vary a lot between professionals in some situations [47].

A useful remark comes from [48]: “Such solutions can reduce the burden on medical professionals and increase patient satisfaction.” This is in line with the following review on Product Hunt [49] about this prototype: “Talk about making doctor visits a little more fun and less intimidating.”

In fact, that was also the motivation behind [12], from which the datasets and some guidance for the image-triggered model in this paper were obtained. They also highlighted the importance of having those systems in place where access to specialized health care professionals is limited.

It is true that we should be cautious about letting these models work without human assistance, but the true question is the scenarios where no human assistance exists at all. In those scenarios, these systems may be an alternative. If no assistance is possible due to the diagnosis being too specialized and expensive, a model could make the difference. Reducing costs in medicine can be a matter worth considering when deciding to deploy those models [50]. I do agree with [38] that chatbots will never replace medical doctors; instead, they can be a first contact, a triage tool, or a health care professional’s assistant.

Furthermore, as demonstrated, a chatbot powered by OpenAI APIs can answer questions using the extensive knowledge acquired during LLM training [51], and such models are increasingly being deployed in medical settings [7-9].

Another remark worth mentioning comes from [52]: “Users should be vigilant of existing chatbots’ limitations, such as misinformation, inconsistencies, and lack of human-like reasoning abilities.” I have shown an example in [4] where the chatbot, which uses the same methodology explored herein, created an entire argumentation to support a wrong prediction, which resulted from the wrong function calling. Wrong function calling is something to pay attention to since it may induce wrong conclusions and misinformation from chatbots. LLMs do not seem to be good at reasoning (eg, spotting wrong vs true argumentation).

There are two possible solutions to misinformation coming from these LLMs: fine-tuning the models from OpenAI or using medical text datasets, which can be articles. The OpenAI API has been shown to be very good at mining information from piles of texts. I have followed a different approach, which can be integrated with these mentioned approaches in the future; they are not incompatible. I have provided functions and trained models that the chatbot can use at their will. This was done using the APIs from OpenAI. This same approach was used by Wolfram Group [10], where they handled the well-known undesirable behavior of ChatGPT to produce disinformation by providing models that could be used for answering questions. There is a growing body of research assessing the place of LLMs in medicine [51]. More discussion of these topics can be found in previously published papers [40-46].

The image-based models were powered by transfer learning. Transfer learning reduces the number of images needed to train the models, the computational demands, and the time needed to converge the models. Thus, it is a widely used approach nowadays to create image-based models [12,28,53-60].

My focus herein is pneumonia detection using X-ray images; therefore, pneumonia works are more related to the discussed endeavor. A recent *Business Insider* article [61] explored how generative AI is being integrated into radiology to automate report writing and facilitate communication.

Regarding related approaches, [62] used ResNet50V2 instead of the classic MobileNet (which was used here), as did the main reference for this work [12]. ResNet50V2 and MobileNet are both convolutional neural networks that are widely used in computer vision tasks.

Several studies highlight the importance of diagnosing COVID-19 pneumonia via chest X-rays [63-65], as early detection can prevent complications like ventilator-associated pneumonia [66]; although COVID-19 often leads to viral pneumonia, which may be less severe than bacterial forms [12], existing systems could be adapted with specific triggers to distinguish COVID-19–related cases.

The main reference [12] related to this work used a similar technique to the one used here. They applied transfer learning using ImageNet for classifying human OCT images. They compared this with human experts and found that even though those models were not better than all experts, they were better than some of the experts. The most interesting result was seeing that those models had less variation within their diagnoses; they tended to be more reliable and predictable with their OCT diagnosis.

Shifting the discussion to the text-based pathway of the chatbot, which handles text input, neural networks have been widely used in the detection and diagnosis of diabetes [67-70]. This paper showcased the text-based capability of the chatbot on a neural network–based diabetes model, which uses physiological measures to make a prediction. The chatbot’s text-triggered pathway autonomously extracts user-provided information and prepares it for the model, enabling fully automated processing.

The focus was on a shallow neural network with no transfer learning and a small number of layers and neurons. The model used here (and possible variations) is from a previous work [13]. Transfer learning is commonly used for image-based models.

Strengths and Limitations

A feature of the currently implemented design is that the machine learning (ie, "the brain") and the app (ie, the chatbot) are decoupled. In practical terms, it is possible to work on them independently. This means that if the approach receives massive investment, the teams can work almost independently.

The models from TM are deployed on their server at Google at no charge from Google's side. When the model is updated or upgraded, the changes will automatically be pushed to the app, even when one adds new classes. It also includes apps from other researchers that may eventually use the models (the models are available as links and can be requested as JSON files). It is in line with a comparative mindset, common in open-source projects.

The Angular app (ie, the chatbot) was deployed on Heroku, a paid server, but it can be deployed using any server service, such as Amazon Web Services. Heroku was chosen due to being very friendly toward Node.js and all the technologies that revolve around it. It is very easy and straightforward to deploy such apps in Heroku. In addition, Heroku has a monthly payment that is independent of the number of apps deployed, so a single account can deploy several apps. There are pricing plans designed for different project stages "from personal projects to enterprise applications."

There are several free web-based medical datasets (eg, on Kaggle). This is ideal for the current system since new models can be added with time, making it smarter and smarter. To add new models and increase the number of possible diagnoses, one just needs to create a model on TM and make the link available. With extra coding, it is also possible to use models created outside TM. It would be possible to create an admin dashboard in the future, where one could just add the link for the model, with no need to make changes to the code.

For the TensorFlow.js models (ie, the text-triggered path), it would be possible to repeat the approach from TM by creating a server just for the models using the link approach. Currently, it is necessary to save the model locally and load it. Those changes could make the platform less dependent on programmers to constantly make changes. Since TM is built on top of TensorFlow.js, it is possible to implement versions of the chatbot that will actually learn instead of just being a hub of pretrained models.

Future Directions

The core usage of function calling is intelligently picking the right model since a trained model will classify anything it is given, even when it makes no sense for the classification task (eg, classifying an X-ray image with an OCT model).

Another option that could serve the same purpose would be a trained model, maybe using just superclasses such as "X-ray images" and "OCT," and then branching out to the right model. It seems that MobileNet can identify X-ray images. These alternatives can replace the use of paid APIs from OpenAI. Open-source LLMs can also be an alternative when considering costs [71].

Recently, Google launched Gemini, which would be interesting to consider as an alternative model to use in future work. I tested it in another study as a chatbot for snake classification, and the results were promising [4].

Current word limits have not been reached yet, but they affect how many functions can be called, since functions are converted to text and count toward the limit. These limits, known as "attention" [72], are being increased and may no longer be an issue in the future. Google's Gemini reportedly supports up to 700,000 words [73], though it is unclear if it also supports function calling.

One issue with fine-tuning models to enhance the chatbot's behavior toward our goals is that there is a cost for this fine-tuning, and the final model costs more than the standard model. This can lead to an increase in the cost of the app. The current limit seems high enough. As one example, *gpt-3.5-turbo-1106* has a limit of 16,385 tokens (about 12,000 words or 50 pages); the GPT-4 model used here has a limit of 128,000 tokens. Those numbers seem to be more than enough, at least for an initial system.

One observation regarding the current prototype is that, currently, even though both the image- and text-based path are triggered using the same interface, they are not aware of each other. It would be interesting to study ways to properly integrate them.

Potential Dangers and Ethical Implications

One possible risk when using chatbots for real-world scenarios, which is significant to mention, is that it is well known that it is not possible to predict with certainty the output from those chatbots (LLMs) [74,75]. Generally, the outputs and performance are within expected behaviors; in the case of OpenAI's API, they are constantly working to increase predictability and moderation. Nonetheless, this is a risk that should be considered when chatbots are left to their own devices [76].

Deploying chatbots in real-world health care scenarios brings both benefits and risks. However, these risks are not necessarily greater than those posed by human professionals. Medical errors are not uncommon and can be severe, depending on the diagnosis. Like human experts, AI models can also make mistakes. Although it is crucial to acknowledge these risks, it is equally important to avoid hasty generalizations. Although cognitive errors in humans (so-called clinical judgment) are well-studied, our understanding of machine errors (so-called mechanical judgment) is still developing [16].

The discussion of the potential dangers and ethical implications of using this chatbot in a real-scenario goes beyond the scope of this paper.

One possible danger is when a model makes a mistake, which they do at times (see the Results section). The model may misclassify urgent and nonurgent pneumonia cases (type I and II errors). To mitigate this, it can be trained to favor safer errors, like classifying nonurgent cases as urgent. Still, mistakes and accountability remain concerns. These models should assist,

not replace, medical professionals. Clear warnings—like those used by OpenAI—should be included to prevent misuse.

This research suggests that chatbots may reduce the risk of model misuse, as users never directly access the models. Errors can occur without domain expertise—for example, misinterpreting model outputs as probabilities. Chatbots help by selecting models, interpreting outputs, and delivering user-friendly responses. As shown in [3], they can even infer implicit information from medical data.

Conclusions

In this paper, I have presented a prototype for a medical chatbot that integrates several models. Integration is a common challenge faced in bioinformatics.

Models are typically developed by separate research groups and published independently, making them hard to integrate into larger integrative frameworks. Models are often built in different languages and formats, without ready-to-use interfaces like APIs or JSON, limiting integration into larger systems. However, I found that TM models can replicate most computer vision tasks without requiring a deep understanding of code. Given the low reproducibility in the field of bioinformatics, it is significant that these models only need training images and basic instructions, as demonstrated with [12].

Tools like TensorFlow.js have made such transfer learning integrations more accessible. Since the LLMs from OpenAI gained momentum, there has been a global race centered around developing increasingly powerful LLMs that resemble or aim toward artificial general intelligence. This is beneficial for bioinformatics, as can be seen in this study. It means that one does not have to build an LLM to make a chatbot that makes their models more user-friendly for their potential users (eg, medical doctors and biologists). This means that UI/UX may actually change in the future—instead of interfaces, we may have chatbots.

A search of the literature for chatbots similar to the one described here showed a considerable increase after the release of OpenAI's LLMs, with most applications concentrated in medicine. The future of AI lies in public APIs—AI as a service—as demonstrated by the feasibility of building complex models without costly research infrastructure. This approach may offer the field of bioinformatics a new ally in the development of more user-friendly interfaces [14].

Acknowledgments

Generative artificial intelligence (AI) tools were used as supportive resources during the preparation of this manuscript. All text was written by JGP, with AI outputs treated strictly as suggestions. Specific uses include (1) spelling and grammar checks, (2) a literature search via RefWiz, which uses the OpenAI application programming interface, and (3) initial literature exploration using Bing Chat, which contributed to early drafts of the Introduction section—these portions were substantially revised before inclusion. Writefull, an AI-powered editing tool, was used to identify textual improvements. Additional minor uses of AI not deemed significant for disclosure were also involved.

Data Availability

All the data used for this manuscript are publicly available. Links are provided in the Multimedia Appendices.

Authors' Contributions

JGP is the sole author of this paper and contributed as follows, according to CRediT taxonomy: conceptualization, investigation, writing—original draft, writing—review & editing, software, validation, and methodology.

Conflicts of Interest

None declared.

Multimedia Appendix 1
The statistics for the models under the hood (the models used by the chatbot).
[PDF File, 428 KB - xmed_v6i1e56090_app1.pdf]

Multimedia Appendix 2
Repository of complete conversations with the chatbot.
[PDF File, 862 KB - xmed_v6i1e56090_app2.pdf]

References

1. Singh RP, Hom GL, Abramoff MD, Campbell JP, Chiang MF, on behalf of the AAO Task Force on Artificial Intelligence. Current challenges and barriers to real-world artificial intelligence adoption for the healthcare system, provider, and the patient. Trans Vis Sci Tech 2020 Jan 28;9(2):45. [doi: 10.1167/tvst.9.2.45]
2. Pires JG. An informal survey presents the gap between computer and medical doctors and biologists. Theoretical and Mathematical Biology (Medium Blog). 2021. URL: https://medium.com/theoretical-and-mathematical-biology/

an-informal-survey-presents-the-gap-between-computer-and-medical-doctors-and-biologists-ca8816051739 [accessed 2025-09-26]
3. Pires JG. Data science using openai: testing their new capabilities focused on data science. Qeios. Preprint posted online on Jan 29, 2024. [doi: 10.32388/76QMHB.2]
4. Pires JG. SnakeChat: a conversational-AI based app for snake classification. Qeios. Preprint posted online on Nov 23, 2023. [doi: 10.32388/Y13B20]
5. Khan HR, Haura I, Uddin R. RoboDoc: smart robot design dealing with contagious patients for essential vitals amid COVID-19 pandemic. Sustainability 2023;15(2):1647. [doi: 10.3390/su15021647]
6. Edmond EC, Prakash E, Carroll F. RoboDoc: critical ethical issues to consider for the design and development of a robotic doctor experience. FRL 2020;1(1):59-63. [doi: 10.3233/FRL-200002]
7. Laranjo L, Dunn AG, Tong HL, et al. Conversational agents in healthcare: a systematic review. J Am Med Inform Assoc 2018 Sep 1;25(9):1248-1258. [doi: 10.1093/jamia/ocy072] [Medline: 30010941]
8. Kell G, Roberts A, Umansky S, et al. Question answering systems for health professionals at the point of care-a systematic review. J Am Med Inform Assoc 2024 Apr 3;31(4):1009-1024. [doi: 10.1093/jamia/ocae015] [Medline: 38366879]
9. Şişman A, Acar AH. Artificial intelligence-based chatbot assistance in clinical decision-making for medically complex patients in oral surgery: a comparative study. BMC Oral Health 2025 Mar 7;25(1):351. [doi: 10.1186/s12903-025-05732-w] [Medline: 40055745]
10. Wolfram S. ChatGPT gets its “Wolfram superpowers”. Stephen Wolfram. 2023 Mar 23. URL: https://writings.stephenwolfram.com/2023/03/chatgpt-gets-its-wolfram-superpowers/ [accessed 2025-09-26]
11. McGinley L. AI hasn’t killed radiology, but it is changing it. The Washington Post. 2025 Apr. URL: https://www.washingtonpost.com/health/2025/04/05/ai-machine-learning-radiology-software/?utm_source=chatgpt.com [accessed 2025-09-26]
12. Kermany DS, Goldbaum M, Cai W, et al. Identifying medical diagnoses and treatable diseases by image-based deep learning. Cell 2018 Feb 22;172(5):1122-1131. [doi: 10.1016/j.cell.2018.02.010] [Medline: 29474911]
13. Pires JG. Machine learning in medicine using JavaScript: building web apps using TensorFlow.js for interpreting biomedical datasets. medRxiv. Preprint posted online on Dec 21, 2023. [doi: 10.1101/2023.06.21.23291717]
14. Pires JG. Innovating with Biomathematics: the challenge of building user-friendly interfaces for computational biology. Academia Letters 2022. [doi: 10.20935/AL5792]
15. Health informatics. Wikipedia. 2024. URL: https://en.wikipedia.org/wiki/Health_informatics [accessed 2024-03-02]
16. Kahneman D, Sibony O, Sunstein CR. Noise: Little, Brown Spark; 2021.
17. Meehl PE. Clinical Versus Statistical Prediction: A Theoretical Analysis and A Review of the Evidence: University of Minnesota Press; 1954. [doi: 10.1037/11281-000]
18. Brown TB, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, et al. Language models are few-shot learners. 2020 Presented at: Proceedings of the 34th International Conference on Neural Information Processing Systems; Dec 6-12, 2020; Red Hook, NY. [doi: 10.5555/3495724.3495883]
19. Esteva A, Kuprel B, Novoa RA, et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature New Biol 2017 Feb 2;542(7639):115-118. [doi: 10.1038/nature21056] [Medline: 28117445]
20. Kaplan J, McCandlish S, Henighan T, Brown TB, Chess B, Child R, et al. Scaling laws for neural language models. . Preprint posted online on Jan 23, 202010.48550/arXiv.2001.08361.
21. OpenAI. Function calling. 2024. URL: https://platform.openai.com/docs/guides/function-calling/function-calling [accessed 2025-09-26]
22. Bookings introduction. URL: https://dev.wix.com/docs/velo/apis/wix-bookings-backend/bookings/introduction [accessed 2025-10-08]
23. Google Calendar API overview. URL: https://developers.google.com/workspace/calendar/api/guides/overview [accessed 2025-10-08]
24. Heroku | Robodoc. URL: https://robodoc.herokuapp.com/#/tools/diabetes [accessed 2025-10-08]
25. Pires JG, da Silva GF, Weyssow T, et al. Galaxy and MEAN stack to create a user-friendly workflow for the rational optimization of cancer chemotherapy. Front Genet 2021;12:624259. [doi: 10.3389/fgene.2021.624259] [Medline: 33679888]
26. Pires JG. What are the challenges that you will face when using JavaScript libraries in Angular. JavaScript in Plain English. 2021. URL: https://javascript.plainenglish.io/what-javascript-has-to-do-with-angular-6fdd45fd30b7 [accessed 2025-09-26]
27. Pires JG. Python, JavaScript is on your snake tail!. IdeaCoding Lab. 2022. URL: https://medium.com/ideacoding-lab/python-javascript-is-on-you-snake-tail-eb1509b7e4cc [accessed 2025-09-26]
28. Pires JG, Dias Braga LH. SnakeFace: a transfer learning based app for snake classification. RBCA 2023;15(3):80-95. [doi: 10.5335/rbca.v15i3.15028]
29. OpenAI API Reference. URL: https://platform.openai.com/docs/api-reference/introduction [accessed 2025-10-08]
30. Angular — What is Angular?. URL: https://angular.dev/overview [accessed 2025-10-08]
31. TensorFlow.js. URL: https://www.tensorflow.org/js [accessed 2025-10-08]
32. Heroku. URL: https://devcenter.heroku.com/categories/reference [accessed 2025-10-08]
33. Jorge Guerra Pires. GitHub. URL: https://github.com/JorgeGuerraPires [accessed 2025-10-08]

34. Jorge Guerra Pires. Amazon. URL: https://www.amazon.com/stores/Jorge-Guerra-Pires/author/B09QJXMG1F?ref=ap_rdr&isDramIntegrated=true&shoppingPortalEnabled=true [accessed 2025-10-08]
35. Jorge Guerra Pires. Udemy. URL: https://www.udemy.com/user/jorge-guerra-pires/?srsltid=AfmBOoorCGgMP0CHEEjzzf29PyCv-dFHhLcGH9-kV-EwriOAR3ua6022 [accessed 2025-10-08]
36. OpenAI | Model optimization. URL: https://platform.openai.com/docs/guides/model-optimization [accessed 2025-10-08]
37. Lubiana T, Lopes R, Medeiros P, et al. Ten quick tips for harnessing the power of ChatGPT in computational biology. PLoS Comput Biol 2023 Aug;19(8):e1011319. [doi: 10.1371/journal.pcbi.1011319] [Medline: 37561669]
38. Altamimi I, Altamimi A, Alhumimidi AS, Altamimi A, Temsah MH. Artificial intelligence (AI) chatbots in medicine: a supplement, not a substitute. Cureus 2023 Jun;15(6):e40922. [doi: 10.7759/cureus.40922] [Medline: 37496532]
39. Elyoseph Z, Hadar-Shoval D, Asraf K, Lvovsky M. ChatGPT outperforms humans in emotional awareness evaluations. Front Psychol 2023;14:1199058. [doi: 10.3389/fpsyg.2023.1199058] [Medline: 37303897]
40. Kim JK, Chua ME, Rickard M, Lorenzo AJ. ChatGPT and large language model (LLM) chatbots: the current state of acceptability and a proposal for guidelines on utilization in academic medicine. J Pediatr Urol 2023 Oct;19(5):598-604. [doi: 10.1016/j.jpurol.2023.05.018] [Medline: 37328321]
41. Li R, Kumar A, Chen JH. How chatbots and large language model artificial intelligence systems will reshape modern medicine: fountain of creativity or Pandora's box? JAMA Intern Med 2023 Jun 1;183(6):596-597. [doi: 10.1001/jamainternmed.2023.1835] [Medline: 37115531]
42. Loh E. ChatGPT and generative AI chatbots: challenges and opportunities for science, medicine and medical leaders. BMJ Lead 2023 May 2;8(1):e000797. [doi: 10.1136/leader-2023-000797] [Medline: 37192124]
43. Galland J. Chatbots and internal medicine: future opportunities and challenges. Rev Med Interne 2023 May;44(5):209-211. [doi: 10.1016/j.revmed.2023.04.001] [Medline: 37127465]
44. Cheong RCT, Pang KP, Unadkat S, et al. Performance of artificial intelligence chatbots in sleep medicine certification board exams: ChatGPT versus Google Bard. Eur Arch Otorhinolaryngol 2024 Apr;281(4):2137-2143. [doi: 10.1007/s00405-023-08381-3] [Medline: 38117307]
45. Greene A, Greene CC, Greene C. Artificial intelligence, chatbots, and the future of medicine. Lancet Oncol 2019 Apr;20(4):481-482. [doi: 10.1016/S1470-2045(19)30142-1] [Medline: 30942174]
46. Miner AS, Laranjo L, Kocaballi AB. Chatbots in the fight against the COVID-19 pandemic. NPJ Digit Med 2020;3:65. [doi: 10.1038/s41746-020-0280-0] [Medline: 32377576]
47. Abimanyi-Ochom J, Bohingamu Mudiyanselage S, Catchpool M, Firipis M, Wanni Arachchige Dona S, Watts JJ. Strategies to reduce diagnostic errors: a systematic review. BMC Med Inform Decis Mak 2019 Aug 30;19(1):174. [doi: 10.1186/s12911-019-0901-1] [Medline: 31470839]
48. Aksenova EI, Medvedeva EI, Kroshilin SV. Chatbots is the modern reality of consulting in medicine. Zdravoohran Ross Fed 2023;67(5):403-410. [doi: 10.47470/0044-197X-2023-67-5-403-410]
49. Product Hunt | Robodoc. URL: https://www.producthunt.com/products/robodoc/launches/robodoc?comment=3125845 [accessed 2025-10-08]
50. Pires JG. Alguns insights em Startups um novo paradigma para a tríplice aliança ciência, tecnologia e inovação [Article in Portuguese]. Rev G&S 2020;11(1):38-54. [doi: 10.26512/gs.v11i1.28626]
51. Rosoł M, Gąsior JS, Łaba J, Korzeniewski K, Młyńczak M. Evaluation of the performance of GPT-3.5 and GPT-4 on the Polish Medical Final Examination. Sci Rep 2023 Nov 22;13(1):20512. [doi: 10.1038/s41598-023-46995-z] [Medline: 37993519]
52. Yang HS, Wang F, Greenblatt MB, Huang SX, Zhang Y. AI chatbots in clinical laboratory medicine: foundations and trends. Clin Chem 2023 Nov 2;69(11):1238-1246. [doi: 10.1093/clinchem/hvad106] [Medline: 37664912]
53. Matsoukas C, Haslum JF, Sorkhei M, Soderberg M, Smith K. What makes transfer learning work for medical images: feature reuse & other factors. Presented at: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); Jun 18-24, 2022; New Orleans, LA. [doi: 10.1109/CVPR52688.2022.00901]
54. Tang H, Cen X. A survey of transfer learning applied in medical image recognition. Presented at: 2021 IEEE International Conference on Advances in Electrical Engineering and Computer Applications (AEECA); Aug 27-28, 2021; Dalian, China p. 94-97. [doi: 10.1109/AEECA52519.2021.9574368]
55. Wang Y. A new classification method for COVID-19 CT images based on transfer learning and attention mechanism. Presented at: 2022 16th ICME International Conference on Complex Medical Engineering (CME); Nov 4-7, 2022; Zhongshan, China p. 236-240. [doi: 10.1109/CME55444.2022.10063276]
56. Dikmen M. Investigating transfer learning performances of deep learning models for classification of GPR B-scan images. TS 2022 Nov 30;39(5):1761-1766. [doi: 10.18280/ts.390534]
57. Aftab MO, Javed Awan M, Khalid S, Javed R, Shabir H. Executing spark BigDL for leukemia detection from microscopic images using transfer learning. Presented at: 2021 1st International Conference on Artificial Intelligence and Data Analytics (CAIDA); Apr 6-7, 2021; Riyadh, Saudi Arabia. [doi: 10.1109/CAIDA51941.2021.9425264]
58. Mahanty C, Kumar R, Mishra BK, Barna C, Balas VE. COVID-19 detection with X-ray images by using transfer learning. Journal of Intelligent & Fuzzy Systems 2022 Jun 9;43(2):1717-1726. [doi: 10.3233/JIFS-219273]

59. Polat Ö, Güngen C. Classification of brain tumors from MR images using deep transfer learning. J Supercomput 2021 Jul;77(7):7236-7252. [doi: 10.1007/s11227-020-03572-9]
60. Yang D, Martinez C, Visuña L, Khandhar H, Bhatt C, Carretero J. Detection and analysis of COVID-19 in medical images using deep learning techniques. Sci Rep 2021;11(1). [doi: 10.1038/s41598-021-99015-3]
61. Khan A. Radiology is embracing generative AI to streamline productivity—and not to replace doctors. Business Insider. 2025 Jun. URL: https://www.businessinsider.com/radiology-embraces-generative-ai-to-streamline-productivity-2025-6 [accessed 2025-09-26]
62. Prusty S, Patnaik S, Dash SK. ResNet50V2: a transfer learning model to predict pneumonia with chest X-ray images. Presented at: 2022 International Conference on Machine Learning, Computer Systems and Security (MLCSS); Aug 5-6, 2022; Bhubaneswar, India. [doi: 10.1109/MLCSS57186.2022.00046]
63. Jawahar M, Anbarasi LJ, Jayachandran P, Ramachandran M, Al-Turjman F. Utilization of transfer learning model in detecting COVID-19 cases from chest X-ray images. International Journal of E-Health and Medical Communications 2022 Jul;13(2):1-11. [doi: 10.4018/IJEHMC.20220701.oa2]
64. Ohata EF, Bezerra GM, Chagas JD, et al. Automatic detection of COVID-19 infection using chest X-ray images through transfer learning. IEEE/CAA J Autom Sinica 2021;8(1):239-248. [doi: 10.1109/JAS.2020.1003393]
65. Hamida S, El Gannour O, Cherradi B, Raihani A, Moujahid H, Ouajji H. A novel COVID-19 diagnosis support system using the stacking approach and transfer learning technique on chest X-ray images. J Healthc Eng 2021;2021:9437538. [doi: 10.1155/2021/9437538] [Medline: 34777739]
66. Deng J, Li F, Zhang N, Zhong Y. Prevention and treatment of ventilator-associated pneumonia in COVID-19. Front Pharmacol 2022;13:945892. [doi: 10.3389/fphar.2022.945892] [Medline: 36339583]
67. ACO-based type 2 diabetes detection using artificial neural networks. Indian Journal of Forensic Medicine & Toxicology 2020;15(1):1765-1771. [doi: 10.37506/ijfmt.v15i1.13666]
68. Haritha R, Sureshbabu D, Sammulal P. Diabetes detection using principal component analysis and neural networks. In: Santosh KC, Hegadi RS, editors. Recent Trends in Image Processing and Pattern Recognition: Springer Singapore; 2019:270-285. [doi: 10.1007/978-981-13-9184-2_24]
69. Alghamdi HS. Towards explainable deep neural networks for the automatic detection of diabetic retinopathy. Appl Sci (Basel) 2022;12(19):9435. [doi: 10.3390/app12199435]
70. Ahmed DW, Hashim FA, Salem NM. Diabetic retinopathy detection using convolutional neural networks. Presented at: 2023 33rd International Conference on Computer Theory and Applications (ICCTA); Dec 16-18, 2023; Alexandria, Egypt p. 240-245. [doi: 10.1109/ICCTA60978.2023.10969372]
71. Bai J, Kamatchinathan S, Kundu DJ, Bandla C, Vizcaíno JA, Perez-Riverol Y. Open-source large language models in action: a bioinformatics chatbot for PRIDE database. Proteomics 2024 Nov;24(21-22):e2400005. [doi: 10.1002/pmic.202400005] [Medline: 38556628]
72. Stern J. GPT-4 has the memory of a goldfish. The Atlantic. 2023. URL: https://www.theatlantic.com/technology/archive/2023/03/gpt-4-has-memory-context-window/673426/ [accessed 2025-09-26]
73. Raveendran C. Gemini 15 - breaking the token limits to create a new future. LinkedIn. 2024. URL: https://www.linkedin.com/pulse/gemini-15-breaking-token-limits-create-new-future-raveendran-i4jqc [accessed 2025-09-26]
74. Zhao H, Chen H, Yang F, Liu N, Deng H, Cai H, et al. Explainability for large language models: a survey. arXiv. Preprint posted online on Sep 2, 2023. [doi: 10.48550/arXiv.2309.01029]
75. Weidinger L, Uesato J, Rauh M, et al. Taxonomy of risks posed by language models. 2022 Jun 21 Presented at: FAccT '22; Jun 21-24, 2022; Seoul, Republic of Korea p. 214-229. [doi: 10.1145/3531146.3533088]
76. Mesinovic M, Watkinson P, Zhu T. Explainability in the age of large language models for healthcare. Commun Eng 2025 Jul 17;4(1):128. [doi: 10.1038/s44172-025-00453-y] [Medline: 40676176]

Abbreviations

API: application programming interface
LLM: large language model
OCT: optical coherence tomography
TM: Teachable Machine
UI/UX: user interface/user experience

Edited by T Leung; submitted 05.01.24; peer-reviewed by Anonymous, Anonymous; revised version received 17.07.25; accepted 28.08.25; published 15.10.25.

Please cite as:
Pires JG
Development of a Conversational Artificial Intelligence–Based Web Application for Medical Consultations: Prototype Study
JMIRx Med 2025;6:e56090
URL: https://xmed.jmir.org/2025/1/e56090
doi:10.2196/56090

Assessing the Limitations of Large Language Models in Clinical Practice Guideline–Concordant Treatment Decision-Making on Real-World Data: Retrospective Study

Tobias Roeschl[1,2,3,4,5*], MD; Marie Hoffmann[2,4,5*], PhD; Djawid Hashemi[1,2,3,4], MD, PD; Felix Rarreck[2,5]; Nils Hinrichs[2,4,5], MSc; Tobias Daniel Trippel[1,2,4], MD, Prof Dr Med; Matthias I Gröschel[2,6], MD, PhD; Axel Unbehaun[2,5], MD, PD; Christoph Klein[2,5], MD, PD; Jörg Kempfert[2,5], MD, Prof Dr Med; Henryk Dreger[1,2], MD, Prof Dr Med; Benjamin O'Brien[2,7,8], MD, Prof Dr Med; Gerhard Hindricks[1,2], MD, Prof Dr Med; Felix Balzer[2,9], MD, PhD, Prof Dr Med; Volkmar Falk[2,4,5,10], MD, Prof Dr Med; Alexander Meyer[2,4,5,11], MD, Prof Dr Med

[1]Department of Cardiology, Angiology and Intensive Care Medicine, Deutsches Herzzentrum der Charité, Berlin, Germany

[10]Department of Health Sciences and Technology, Translational Cardiovascular Technologies, Institute of Translational Medicine, Swiss Federal Institute of Technology, Zürich, Switzerland

[11]Berlin Institute for the Foundations of Learning and Data – TU Berlin, Berlin, Germany

[2]Charité – Universitätsmedizin Berlin, corporate member of Freie Universität Berlin and Humboldt-Universität zu Berlin, Charitéplatz 1, Berlin, Germany

[3]Berlin Institute of Health at Charité – Universitätsmedizin Berlin, BIH Biomedical Innovation Academy, BIH Charité Digital Clinician Scientist Program, Berlin, Germany

[4]DZHK (German Centre for Cardiovascular Research), partner site Berlin, Berlin, Germany

[5]Department of Cardiothoracic and Vascular Surgery, Deutsches Herzzentrum der Charité (DHZC), Berlin, Germany

[6]Department of Infectious Diseases and Respiratory Medicine, Charité – Universitätsmedizin Berlin, Berlin, Germany

[7]Department of Cardiac Anesthesiology and Intensive Care Medicine, Deutsches Herzzentrum der Charité (DHZC), Berlin, Germany

[8]Department of Perioperative Medicine, St Bartholomew's Hospital and Barts Heart Centre, London, United Kingdom

[9]Charité – Universitätsmedizin Berlin, Institute of Medical Informatics, Berlin, Germany

[*]these authors contributed equally

Corresponding Author:
Marie Hoffmann, PhD
Charité – Universitätsmedizin Berlin, corporate member of Freie Universität Berlin and Humboldt-Universität zu Berlin, Charitéplatz 1, Berlin, Germany

Related Articles:

Companion article: http://preprints.jmir.org/preprint/74899

Companion article: https://med.jmirx.org/2025/1/e84175

Companion article: https://med.jmirx.org/2025/1/e84174

Companion article: https://med.jmirx.org/2025/1/e84173

Abstract

Background: Studies have shown that large language models (LLMs) are promising in therapeutic decision-making, with findings comparable to those of medical experts, but these studies used highly curated patient data.

Objective: This study aimed to determine if LLMs can make guideline-concordant treatment decisions based on patient data as typically present in clinical practice (lengthy, unstructured medical text).

Methods: We conducted a retrospective study of 80 patients with severe aortic stenosis who were scheduled for either surgical (SAVR; n=24) or transcatheter aortic valve replacement (TAVR; n=56) by our institutional heart team in 2022. Various LLMs (BioGPT, GPT-3.5, GPT-4, GPT-4 Turbo, GPT-4o, LLaMA-2, Mistral, PaLM 2, and DeepSeek-R1) were queried using either anonymized original medical reports or manually generated case summaries to determine the most guideline-concordant treatment. We measured agreement with the heart team using Cohen κ coefficients, reliability using intraclass correlation coefficients (ICCs), and fairness using the frequency bias index (FBI; FBI >1 indicated bias toward TAVR).

Results: When presented with original medical reports, LLMs showed poor performance (Cohen κ coefficient: −0.47 to 0.22; ICC: 0.0 - 1.0; FBI: 0.95 - 1.51). The LLMs' performance improved substantially when case summaries were used as input and additional guideline knowledge was added to the prompt (Cohen κ coefficient: −0.02 to 0.63; ICC: 0.01 - 1.0; FBI: 0.46 - 1.23). Qualitative analysis revealed instances of hallucinations in all LLMs tested.

Conclusions: Even advanced LLMs require extensively curated input for informed treatment decisions. Unreliable responses, bias, and hallucinations pose significant health risks and highlight the need for caution in applying LLMs to real-world clinical decision-making.

(JMIRx Med 2025;6:e74899) doi:10.2196/74899

KEYWORDS

large language models; foundation models; reasoning models; treatment decision-making; aortic stenosis; clinical practice guidelines; medical data processing

Introduction

Large language models (LLMs) have recently demonstrated their impressive capabilities in medicine, exemplified by passing medical board exams [1], making correct diagnoses in complex clinical cases [2], and excelling in physician-patient communication [3]. Most recently, the use of LLMs in therapeutic decision-making has been trialed. Several studies have shown that LLMs can make treatment decisions for patients with oncological and cardiovascular diseases that are in substantial agreement with the respective treatment decisions made by clinical experts on tumor boards [4-7] and heart teams (HTs) [8]. However, a common feature of these studies was that the LLMs did not make treatment decisions based on real-world patient data in its original format (eg, discharge letters, imaging reports, etc) but rather made decisions based on preprocessed data.

In clinical practice, relevant patient data, such as patient characteristics, comorbidities, tumor stages, and imaging results, are typically available in free-text format, either as medical text reports or as text entries in the electronic health record, a format that is likely to persist in the near future. In the aforementioned studies, however, decision-relevant patient data were extracted from the original medical reports by the investigators in a preprocessing step before being provided to the LLMs as input in a concise and high-quality form. However, it is still unknown to what extent LLMs can make treatment decisions based on the original medical data, a scenario that could lead to a significant reduction in physician workload and potentially increase guideline adherence and thus improve patient care.

In this study, we investigated the impact of data representation, that is, using original medical reports versus case summaries, on the performance of LLMs in therapeutic decision-making.

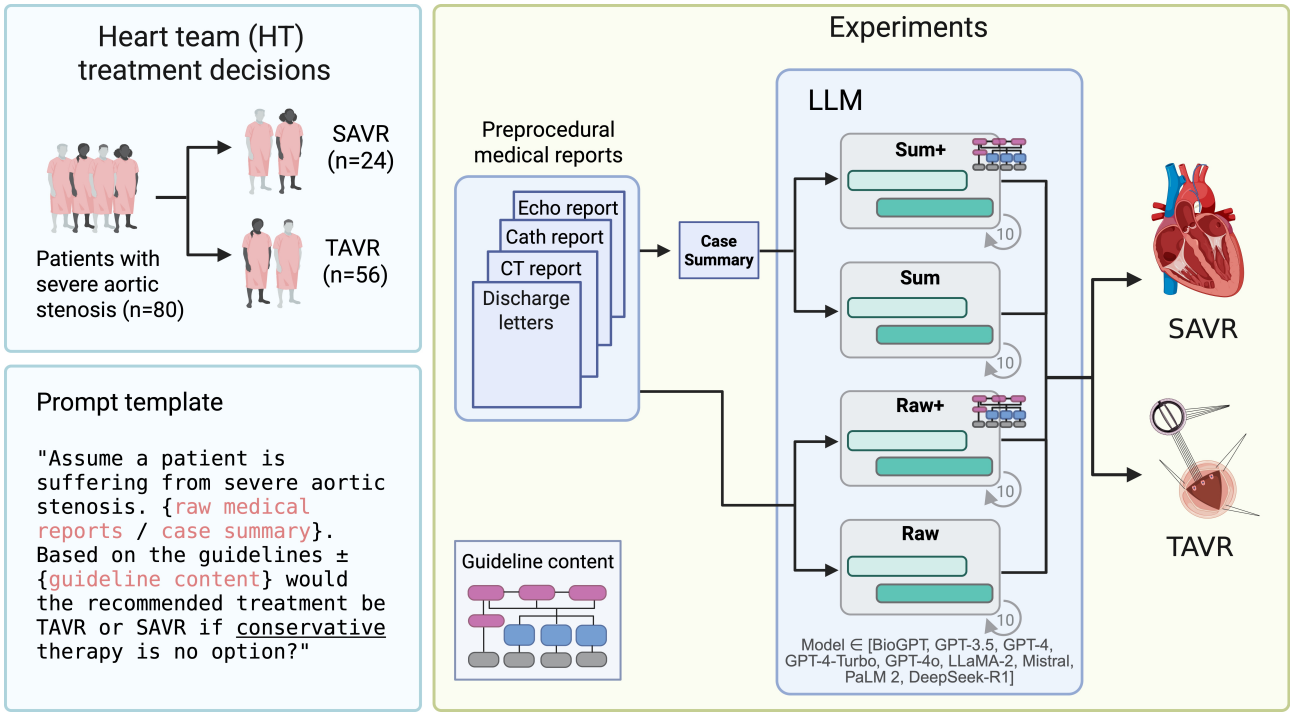
As our study population, we selected patients with severe aortic stenosis (AS). This cohort was chosen because the parameters relevant to decision-making are readily quantifiable, the potential for resource optimization is substantial, and the prevalence of the condition is increasing. If left untreated, AS is associated with high morbidity and mortality [9]. Treatment modalities for severe AS include surgical aortic valve replacement (SAVR), transcatheter aortic valve replacement (TAVR), and, to a lesser extent, medical therapy. The choice of the optimal treatment modality depends on several clinical variables, including patient age, estimated surgical risk, comorbidities, and anatomical factors, as specified in the 2021 European Society of Cardiology (ESC) and European Association for Cardio-Thoracic Surgery (EACTS) Guidelines for the management of valvular heart disease [10]. The 2021 ESC/EACTS Guidelines strongly endorse an active, collaborative consultation with a multidisciplinary HT. HTs are comprised of cardiologists, cardiac surgeons, cardiac imaging specialists, and cardiac anesthesiologists. In HT meetings, these experts review a patient's condition based on patient data laboriously extracted from medical reports before arriving at a treatment decision using a guideline-based approach.

Methods

Study Design and Evaluation Framework

We presented patient data to an LLM to obtain a treatment decision of either SAVR or TAVR. We assessed the degree of agreement between the treatment decisions provided by the LLM and the treatment decisions provided by the HT. Furthermore, we assessed the decidability, reliability, and fairness of the LLM. Finally, we compared the performance of 7 state-of-the-art LLMs to the performance of a simple non-LLM reference model. In an ablative manner, we studied the effect of using case summaries instead of the original medical reports and adding guideline knowledge to the prompt separately, resulting in 4 distinct experiments (Figure 1).

Figure 1. Experimental design. We presented the clinical data of 80 patients with severe aortic stenosis to a large language model (LLM) to receive a treatment decision for either surgical aortic valve replacement (SAVR) or transcatheter aortic valve replacement (TAVR), repeating each query 10 times. To investigate whether injecting guideline knowledge (raw+) into the prompt and/or using case summaries (sum and sum+) instead of the original medical reports (raw) improves LLM performance, we conducted a total of 4 experiments. Case summaries included only decision-relevant patient data and were manually created by physicians. CT: computed tomography.



Study Population

This study included patients treated at a heart center. We screened all patients with severe degenerative AS who were scheduled for an HT meeting in our hospital information system at 1 campus of our center in 2022. We identified 80 patients with sufficiently digitized documentation. As part of a quaternary care center, our institutional HT receives preselected patients scheduled for invasive AS treatment. Therefore, the number of patients recommended for conservative treatment at our institution is negligible. As a result, we decided to limit the possible therapeutic options for this study to SAVR and TAVR, excluding conservative therapy.

Ethical Considerations

This study was approved by the research ethics committee of Charité – Universitätsmedizin Berlin (EA1/146/23). The approval included the collection of data based on implied consent owing to the retrospective and observational nature of the study.

Data Collection

Medical reports were available as PDF files in our hospital information system. For each patient, we included the following preprocedural reports: the 2 most recent discharge letters (including letters from external clinics), invasive coronary angiography report, echocardiography report, computed tomography (CT) scan report, and HT report. We manually anonymized these reports prior to analysis.

HT meeting protocols are standardized documents that contain decision-relevant patient characteristics, such as comorbidities, surgical risk scores, and the final treatment decision of the HT (Multimedia Appendix 1). A detailed description of our institutional HT is provided in Figure S1 in Multimedia Appendix 1.

LLMs Assessed

The study used several state-of-the-art LLMs, namely GPT-3.5 [11], GPT-4 [12], GPT-4 Turbo, and GPT-4o by OpenAI, and PaLM 2 by Google [13]. In addition, we used the open-source models DeepSeek-R1 [14] by DeepSeek, Mistral-7B [15], LLaMA-2 by Meta [16], and BioGPT [17]. These LLMs had either demonstrated proficiency in similar tasks or had undergone pretraining on medical literature. Model details are provided in Multimedia Appendix 1. The model hyperparameters were set to the default values, except for the temperature, which was set to zero in accordance with previous studies in the medical domain [18]. Temperature is a hyperparameter that controls the randomness of the LLM's output. Lower values make the output more deterministic and focused, reducing variability and creativity. A detailed description of how we accessed the LLMs and handled input size constraints is given in Multimedia Appendix 1.

Reference Model

The reference model represented an algorithmic emulation of the 2021 ESC/EACTS Guidelines for the management of valvular heart disease [10]. More specifically, the reference model assigned patients to either SAVR or TAVR according to a flowchart (Figure S2 in Multimedia Appendix 1) and relevant clinical variables (Tables S4 and S5 in Multimedia Appendix 1) [10]. Model details are provided in Table S1 in Multimedia Appendix 1.

Experiments

Four experiments were conducted to investigate the effects of data preprocessing on LLM performance: raw, raw+, sum, and sum+.

Raw

In the raw experiment, we programmatically extracted the text content from the PDF files of relevant medical reports (ie, the 2 most recent discharge letters, invasive coronary angiography report, echocardiography report, and CT scan report) using Tesseract and concatenated these into a unified plain-text file. This text file was then manually anonymized and programmatically inserted into a prompt template. Each prompt included an introductory or continuation phrase and concluded with a request for a treatment decision (Table S6 in Multimedia Appendix 1).

Raw+

As it is unknown whether the LLMs we used had sufficient knowledge of clinical practice guidelines (CPGs), we compiled a summary of relevant CPG content from the ESC/EACTS Guidelines [10]. We added this summary to the prompt along with the unified text reports.

Sum

To study the effect of content compression, we replaced the original medical reports used in the raw experiment with concise case summaries. These case summaries were created manually by the study team following a predefined template, with each patient characteristic documented in the HT protocol (Figure S1 in Multimedia Appendix 1) either affirmed, negated, or populated with the patient-specific value, as exemplified in Table S6 in Multimedia Appendix 1.

Sum+

Case summaries were used as input and were enriched with the CPG summary (Figure 1).

Prompt templates, the CPG summary, and an exemplary case summary are shown in Table S6 in Multimedia Appendix 1.

The LLMs' responses were manually reviewed and categorized as either "TAVR," "SAVR," or "indeterminate." Indeterminate responses occur when the model output does not match the available answer choices or when the model determines that there is insufficient information to support a decision (Table S7 in Multimedia Appendix 1). To assess reliability and obtain robust estimates of performance metrics, the LLMs were presented with the same prompt input 10 times in succession for each experiment and patient (hereafter referred to as "runs") to obtain a treatment decision. To prevent memory bias, a new chat session was initiated for each run.

Performance Metrics

We quantified agreement by means of Cohen κ coefficients. For the sake of completeness, we also calculated accuracies as the proportion of treatment decisions that agreed with those made by the HT; however, we emphasize that due to class imbalance, this metric is only of limited significance and therefore only reported in Table S9 in Multimedia Appendix 1. Decidability was quantified as the proportion of determinate treatment decisions. Bias was quantified using the frequency bias index (FBI), defined as the ratio of predicted to observed treatment decisions for TAVR.

Due to the limitations of individual metrics, we used 2 different metrics to quantify reliability: intraclass correlation coefficients (ICCs) and normalized Shannon entropy. A detailed description of the performance metrics, including strategies for handling indeterminate responses, is provided in Table S8 in Multimedia Appendix 1.

Statistical Analysis

The characteristics of patients who received SAVR and those who received TAVR were compared using the Student *t*-test for normally distributed continuous variables and the Mann-Whitney *U* test for variables departing from normality. The Shapiro-Wilk test was used to assess normality. The chi-square test was used for binary variables, and the Fisher exact test was used for sparse binary data.

Accuracy and Cohen κ were computed with Python's sklearn.metrics package (version 1.2.2). ICCs were calculated based on a 1-way random effects, absolute agreement, single-rater model [19] using Python's pingouin package (version 0.5.3).

Results

Patient Characteristics

A total of 80 patients with severe AS who were discussed at our institutional HT in 2022 were included. Of these patients, 24 (30%) underwent SAVR, while 56 (70%) underwent TAVR. Patient characteristics are presented in Table 1.

Table . Patient characteristics.

Variable	Data availability (%)	Overall (N=80)	SAVR[a] (n=24)	TAVR[b] (n=56)	*P* value
Age (years), mean (SD)	100	77.74 (7.5)	70.71 (6.1)	80.75 (5.8)	<.001
Female sex, n (%)	100	36 (45)	8 (33)	28 (50)	.26
Height (cm), mean (SD)	100	168.1 (11.0)	172.5 (11.0)	166.3 (10.6)	.02
Body mass (kg), mean (SD)	100	76.3 (17.0)	79.0 (16.0)	75.1 (17.4)	.35
BMI (kg/m^2), median (IQR)	100	26.0 (23.0-29.7)	25.9 (23.2-29.0)	26.2 (23.0-29.8)	.66
Logistic EuroSCORE[c], median (IQR)	31	6.8 (4.5-13.0)	4.5 (2.2-6.8)	8.4 (5.0-16.0)	.20
EuroSCORE II, median (IQR)	99	2.6 (1.6-4.5)	1.8 (1.1-3.1)	2.9 (1.8-4.9)	.02
STS[d] score, median (IQR)	76	2.8 (1.6-4.5)	1.4 (1.1-3.0)	3.3 (2.1-4.5)	.12
Left ventricular ejection fraction (%), median (IQR)	100	60.0 (54.3-61.3)	60.0 (48.8-62.0)	60.0 (55.0-60.0)	.28
Aortic valve opening area (cm^2), median (IQR)	100	0.70 (0.60-0.80)	0.80 (0.68-0.80)	0.70 (0.60-0.80)	.18
Arterial hypertension, n (%)	100	59 (74)	18 (75)	41 (73)	>.99
Diabetes mellitus, n (%)	100	22 (28)	6 (25)	16 (29)	.96
Hyperlipidemia, n (%)	100	51 (64)	13 (54)	38 (68)	.36
Previous cardiac surgery, n (%)	100	1 (1)	0 (0)	1 (2)	>.99
Frailty, n (%)	100	7 (9)	0 (0)	7 (13)	.17
Sequelae of chest radiation, n (%)	100	0 (0)	0 (0)	0 (0)	>.99
Porcelain aorta, n (%)	100	0 (0)	0 (0)	0 (0)	>.99
Expected patient-prosthesis mismatch, n (%)	100	1 (1)	0 (0)	1 (2)	>.99
Severe chest deformation or scoliosis, n (%)	100	7 (9)	1 (4)	6 (11)	.60
Severe coronary artery disease requiring revascularization, n (%)	100	6 (8)	5 (21)	1 (2)	.01
Left ventricular ejection fraction ≤40%, n (%)	100	6 (8)	3 (13)	3 (5)	.52
Active neoplasia, n (%)	100	7 (9)	2 (8)	5 (9)	>.99
Liver cirrhosis, n (%)	100	1 (1)	0 (0)	1 (2)	>.99
Chronic obstructive pulmonary disease (GOLD[e] stage ≥3), n (%)	100	5 (6)	1 (4)	4 (7)	>.99

Variable	Data availability (%)	Overall (N=80)	SAVR[a] (n=24)	TAVR[b] (n=56)	*P* value
Pulmonary arterial hypertension, n (%)	100	8 (10)	3 (13)	5 (9)	.94
Under immunosuppressive therapy, n (%)	100	10 (13)	2 (8)	8 (14)	.71

[a]SAVR: surgical aortic valve replacement.

[b]TAVR: transcatheter aortic valve replacement.

[c]EuroSCORE: European System for Cardiac Operative Risk Evaluation.

[d]STS: Society of Thoracic Surgeons.

[e]GOLD: Global Initiative for Chronic Obstructive Lung Disease.

Qualitative Analysis

The LLMs' outputs ranged from nonsensical treatment recommendations (eg, heart transplant) and purely fabricated content to correctly assessing the patient's status, choosing the correct treatment option, and supporting the treatment decision with additional anatomical insights (Table 2). Qualitative analysis revealed that smaller models (eg, BioGPT) tended to provide conflicting treatment recommendations for the same patient. In contrast, the frontier models (eg, GPT-4 and PaLM 2) consistently provided the same treatment recommendation when presented with the same patient data repeatedly over 10 runs.

In each experiment, all LLMs produced hallucinations of varying severity and frequency. These included instructional, contextual, and factual inconsistencies (Table 2).

Table . Representative responses from the LLMs[a].

Model	Experiment	Patient characteristics	LLM response[b,c]	HT[d] treatment decision	Interpretation
PaLM 2	Raw	56-year-old male; EuroSCORE[e] II: 0.55%; no comorbidities except diffuse, mild coronary atherosclerosis and arterial hypertension; no relevant anatomical aspects to consider	• Response: "Based on the provided information, the recommended treatment for the patient would be transcatheter aortic valve replacement (TAVR). The patient has severe aortic stenosis, which is causing symptoms and has been confirmed by echocardiography and cardiac catheterization. The patient is also *at high risk for surgical aortic valve replacement (SAVR) due to their age, comorbidities, and the presence of diffuse coronary artery disease*[f]. TAVR is a less invasive procedure than SAVR and is associated with a *lower risk of complications*[f]. Therefore, *TAVR*[f] would be the preferred treatment option for this patient."	SAVR[g]	• The LLM incorrectly recommended TAVR[h] based on falsely extracted and/or interpreted patient characteristics (factual inconsistency). • According to the literature, it has not yet been conclusively determined whether TAVR is associated with fewer complications than SAVR [20] (factual inconsistency/bias).
BioGPT	Raw	69-year-old male; EuroSCORE II: 7.2%; postcardiac arrest syndrome, frailty, long-term mechanical ventilation, and liver cirrhosis; no relevant anatomical aspects to consider	• Response: "*Surgical replacement of the aortic valve. The patient was referred to the operating room for AVR. The operation was performed under general anesthesia and extracorporeal circulation ... The operation was performed with a median sternotomy. The ascending aorta was dissected free and opened…and the coronary arteries were visualized…*[f]"	TAVR	• The LLM falsely stated a surgery report from a procedure this patient never received and that was not mentioned in the prompt (context inconsistency).

Model	Experiment	Patient characteristics	LLM response[b,c]	HT[d] treatment decision	Interpretation
BioGPT	Raw	75-year-old female; EuroSCORE II: 2.4%; STS[i] score: 2.9%; COPD[j] (GOLD[k] Stage III); pulmonary hypertension and frailty; no relevant anatomical aspects to consider	• Response: "*Heart transplantation. - The patient's family decided against surgery. - The patient died due to heart failure.*[f]"	TAVR	• The LLM did not follow the instruction to choose between SAVR and TAVR but opted for heart transplantation (instruction inconsistency). • The LLM fabricated a clinical course that was not mentioned in the prompt (context inconsistency).
GPT-3.5	Sum	72-year-old female; EuroSCORE II: 1.6%; STS score: 1.1%; no relevant comorbidities; no relevant anatomical aspects to consider	• $Response_1$: "Based on the given information, the recommended treatment for a person with severe aortic valve stenosis and no other significant comorbidities would be *SAVR*[l] (Surgical Aortic Valve Replacement) if conservative therapy is not an option." • $Response_2$: "Based on the provided information, the recommended treatment for a person with severe aortic valve stenosis, who does not have any contraindications or relevant comorbidities, would be *TAVI*[f] (Transcatheter Aortic Valve Implantation) rather than SAVR (Surgical Aortic Valve Replacement)."	SAVR	• The LLM gave opposite answers to the exact same question (unreliability). • The LLM incorrectly recommended TAVR based on the absence of comorbidities (factual inconsistency).
PaLM 2	Raw+	56-years-old male; EuroSCORE II: 0.55%; STS score: 0.7%; no comorbidities except arterial hypertension; no relevant anatomical aspects to consider		SAVR	• The LLM falsely claimed that the patient is at high surgical risk due to frailty and previous cardiac surgery, which are both entirely fabricated (factual inconsistency).

Model	Experiment	Patient characteristics	LLM response[b,c]	HT[d] treatment decision	Interpretation
			• Response: "Based on the guidelines, the recommended treatment would be transcatheter aortic valve replacement (*TAVR*[f]). The patient is 56 years old and at *high surgical risk*[f] (STS-PROM score >8% or EuroSCORE II >8 %) due to *severe frailty*[f] and *previous cardiac surgery*[f]. TAVR is the preferred treatment for patients who are inoperable or at high surgical risk."		
GPT-3.5	Sum+	81-year-old female; logistic EuroSCORE: 8.44%; EuroSCORE II: 1.82%; STS score: 4.33%; stage 3A chronic kidney disease; no relevant anatomical aspects to consider	• Response: "Based on the guidelines provided, the recommended treatment for the patient would be *SAVR*[f] (surgical aortic valve replacement) if conservative therapy is not an option. This is because the patient is 81 years old and falls into the *older age category*[l], and *their logistic EuroSCORE*[f] is 8.44%, which is above the threshold for TAVR (transcatheter aortic valve replacement) recommendation."	TAVR	• The LLM incorrectly mentioned an association between the logistic EuroSCORE and surgical risk prohibiting SAVR (context inconsistency). • The LLM correctly assigned the patient to the older age category but then incorrectly recommended SAVR as the optimal therapy (logical inconsistency).
GPT-4	Raw	65-year-old female; EuroSCORE II: 2.5%; STS score: 1.4%; no relevant comorbidities; ascending aortic aneurysm (48 mm) mentioned in the CT[m] scan report		SAVR	• Correct treatment decision plus additional extraction of decision-relevant clinical data.

Model	Experiment	Patient characteristics	LLM response[b,c]	HT[d] treatment decision	Interpretation
			• Response: "Based on the guidelines, the recommended treatment would be Surgical Aortic Valve Replacement (*SAVR*[l]) as the patient has a severe aortic stenosis and an *ascending aortic aneurysm*[l]. Transcatheter Aortic Valve Implantation (TAVI) is typically recommended for patients who are at *high or prohibitive surgical risk, which does not seem to be the case here.*[l]"		
LLaMA-2	Sum+	68-year-old male; EuroSCORE II: 1.29%; STS score: 3.04%; COPD; no relevant anatomical aspects to consider		SAVR	• The LLM correctly identified low surgical risk but then incorrectly rejected SAVR as a therapeutic option (logical inconsistency). • The LLM incorrectly mentioned factors favoring TAVR from the prompt (context inconsistency). • Although the LLM correctly selected SAVR the second time, it gave the opposite response to the exact same question (unreliability).

Model	Experiment	Patient characteristics	LLM response[b,c]	HT[d] treatment decision	Interpretation
			• Reponse$_1$: "Based on the guidelines provided, the recommended treatment for this patient would be TAVI. The patient's age (68 y old) and *low surgical risk*[l] (EuroSCORE II 1.29 %, STS score 3.037 %) make them a suitable candidate for *TAVI*. Additionally, the patient does not have any *contraindications for TAVI*[f], *such as severe frailty, previous cardiac surgery, or significant thorax deformity.*[f]" • Response$_2$: 'Based on the information provided, the patient's EuroSCORE II is 1.29% and STS score is 3.037 %, which indicates that they are at *low surgical risk*[l]. The patient is also *younger than 75 years*[l] old. Therefore, according to the guidelines, *SAVR*[l] is the recommended treatment option for this patient."		
DeepSeek-R1	Raw+	65-year-old male; EuroSCORE II: 0.92%; STS score: 0.73%; end-stage renal disease requiring hemodialysis; no relevant anatomical aspects to consider		SAVR	• The LLM overestimated the effect of end-stage renal disease on surgical risk, which was still relatively low overall, as evidenced by low surgical risk scores (factual inconsistency).

Model	Experiment	Patient characteristics	LLM response[b,c]	HT[d] treatment decision	Interpretation
			• Response: "… Based on the 2021 ESC/EACTS guidelines and the patient's clinical profile, the recommended treatment would be *TAVR*[f]. Reasoning: Patient age (65 y) is below 75 but has significant comorbidities, *including terminal renal insufficiency, indicating high surgical risk*[f]. Suitable tricuspid aortic valve anatomy for TAVR. High surgical risk factors favor TAVR over SAVR."		

[a]LLM: large language model.

[b]The LLMs' treatment responses included well-informed decisions but also hallucinations ranging from obvious misinformation to absurd treatment recommendations and logical errors. We largely adhered to the taxonomy for the description of hallucinations established by Huang et al [21].

[c]LLM responses with subscripts indicate responses to the same question (obtained during 10 runs).

[d]HT: heart team.

[e]EuroSCORE: European System for Cardiac Operative Risk Evaluation.

[f]The italicized part indicates an incorrect or harmful response.

[g]SAVR: surgical aortic valve replacement.

[h]TAVR: transcatheter aortic valve replacement.

[i]STS: Society of Thoracic Surgeons.

[j]COPD: chronic obstructive pulmonary disease.

[k]GOLD: Global Initiative for Chronic Obstructive Lung Disease.

[l]The italicized part indicates a correct or useful response.

[m]CT: computed tomography.

Quantitative Analysis

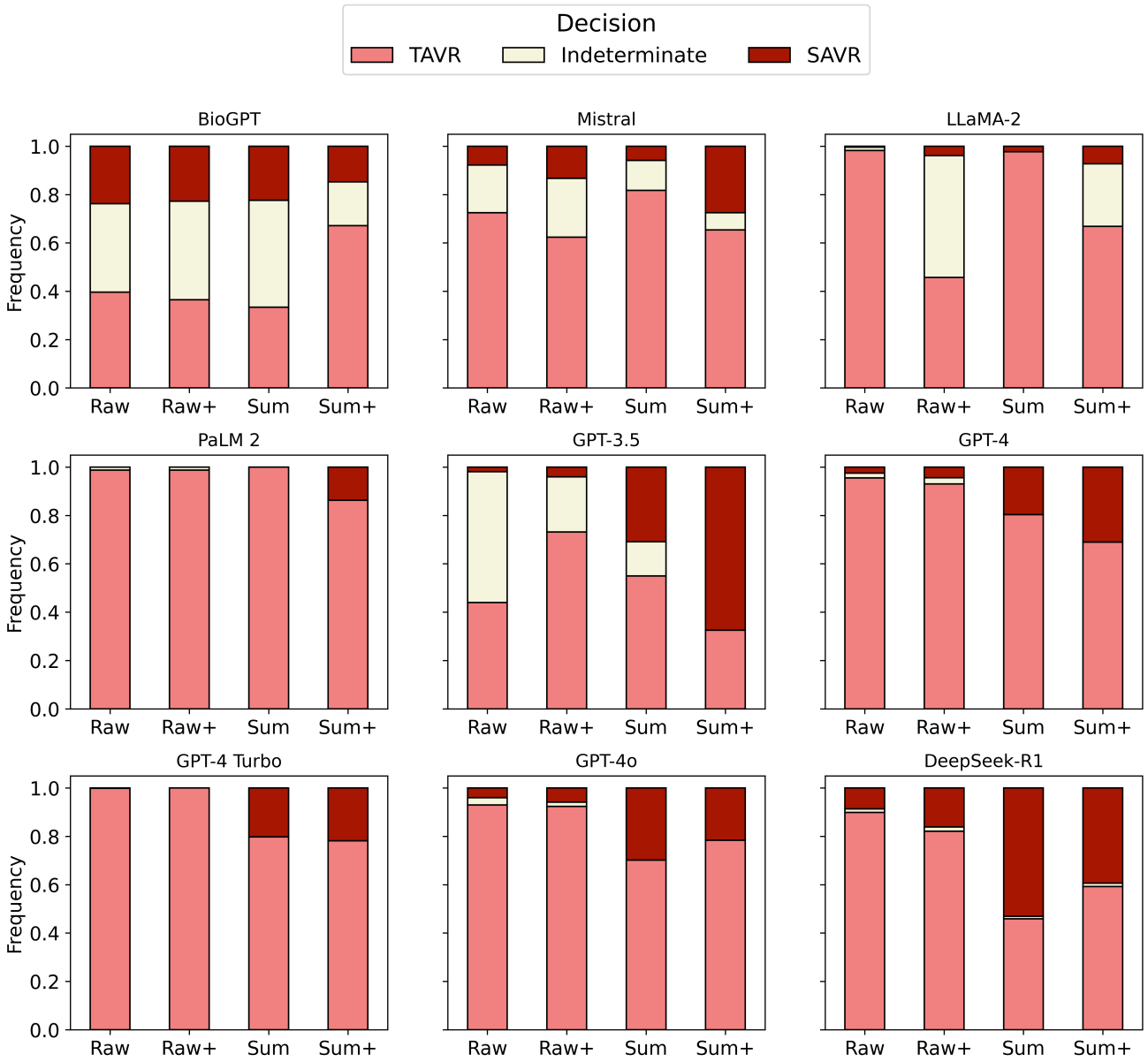
Figure 2 and Table S9 in Multimedia Appendix 1 present the performance metrics. In the raw experiment, LLMs' treatment decisions were in poor agreement with the HT. In this experiment, DeepSeek-R1 showed the highest agreement with the HT, with a Cohen κ coefficient of 0.22. Some LLMs gave indeterminate treatment recommendations in up to 54% of cases (eg, GPT-3.5) and showed low reliability as evidenced by low ICCs and high entropy values (eg, Mistral, LLaMA-2, and DeepSeek-R1). FBIs were substantially higher than 1.0 for all LLMs, except BioGPT, indicating a bias toward TAVR. The reference model outperformed the LLMs in the raw experiment regarding the metrics we assessed.

In the raw+ experiment, DeepSeek-R1 again showed the highest agreement with the HT with a Cohen κ coefficient of 0.40, indicating fair agreement. The performance metrics of the other LLMs did not change substantially in the raw+ experiment. However, the performance metrics of most LLMs substantially improved in the sum experiment and peaked in the sum+ experiment, where some LLMs (eg, GPT-4 models and DeepSeek-R1) drew level with the reference model.

A general trend toward more concordant treatment decisions, fewer indeterminate responses, increased reliability, and less bias toward TAVR was observed with increasing data preprocessing and information enrichment efforts from the raw experiment to the sum+ experiment (Figures 2 and 3).

Figure 2. Performance metrics of the large language models are shown for the 4 experiments conducted. The dashed line represents the reference model. Cohen κ coefficients ≤0 indicate no agreement, 0.01 - 0.20 indicate slight agreement, 0.21 - 0.40 indicate fair agreement, 0.41 - 0.60 indicate moderate agreement, 0.61 - 0.80 indicate substantial agreement, and 0.81 - 1.0 indicate almost perfect agreement [20] with the heart team's treatment decisions. Frequency bias index (FBI) values >1 indicate bias toward transcatheter aortic valve replacement (TAVR) and <1 indicate bias toward surgical aortic valve replacement (SAVR). Intraclass correlation coefficients (ICCs) <0.5 indicate poor test-retest reliability, 0.50 - 0.75 indicate moderate reliability, 0.75 - 0.90 indicate good reliability, and >0.90 indicate excellent reliability [19]. Instances where ICCs were undefined are marked by asterisks. Entropy values close to 0 indicate low output variation, and entropy values close to 1 indicate high output variation. Decidability was defined as the proportion of nonindeterminate treatment decisions. The exact numerical values for the performance metrics are displayed in Table S9 in Multimedia Appendix 1.

Figure 3. Frequencies of the treatment decisions of the large language models in the 4 experiments conducted. A general trend toward increasing decidability and an increasing proportion of treatment decisions favoring surgical aortic valve replacement (SAVR) could be observed between the raw experiment and the sum+ experiment. TAVR: transcatheter aortic valve replacement.



Discussion

LLM Performance With Original Clinical Data

To our knowledge, this is the first study to evaluate the impact of input data representation, including real-world medical data, on the ability of LLMs to make guideline-concordant treatment decisions.

Current LLMs Make Incorrect Decisions Based on Original Clinical Data

Our analysis revealed that LLMs struggled to process original medical reports effectively, often outputting "TAVR" or providing indeterminate responses. The LLMs showed low agreement with the HT, exhibited undecidability and unreliability, and displayed a strong bias toward TAVR. The considerably high accuracies (Table S9 in Multimedia Appendix 1) observed with some LLMs in the raw experiment can be largely attributed to the class imbalance within our patient cohort, where 70% of patients received TAVR.

LLMs Require Extensive Data Preprocessing to Make Sound Therapeutic Decisions

Performance improved substantially when physician-made case summaries were used as input and when guideline knowledge was added to the prompts. The GPT-4 models and DeepSeek-R1 stood out as the most capable LLMs in our experiments. When given case summaries and a CPG summary, these 2 models showed substantial agreement with HT and drew level with the reference model in terms of interrater agreement, decidability, and bias.

Data Representation Affects LLM Performance

GPT-4o, a distilled and streamlined version of GPT-4, and DeepSeek-R1, a model with enhanced reasoning abilities, showed more promising results than previous-generation LLMs when confronted with real-world medical data (raw and raw+

experiments); however, their performances remain insufficient for clinical application. The fact that even state-of-the-art LLMs show significant stochastic variations in decision-making, and thus unreliability, further supports this finding.

An explanation for the underperformance of LLMs in the raw experiment is not immediately apparent due to their opaque nature and a lack of established tools that allow the direct examination of input-output correlations. However, the underperformance cannot be attributed to a lack of guideline knowledge or incorrectly applied guideline knowledge since the performance in the raw+ experiment was, in general, similar to that in the raw experiment and since LLMs can presumably apply clinical knowledge to clinical cases as shown in their ability to pass medical board exams [1,22].

This, along with the significant performance gains observed when providing case summaries instead of original medical reports, suggests that input data representation is the most critical factor in LLM performance. This finding is consistent with the fact that virtually all studies, which showed that LLMs make sound treatment decisions, used preprocessed clinical data as input [4-8]. Of note is the study by Salihu et al [8]. In this study, data from patients with severe AS were provided to GPT-4 to obtain a treatment decision for either TAVR, SAVR, or conservative management. Patient data were provided in the form of a standardized multiple-choice questionnaire with 14 key clinical variables as input, similar to our sum experiments. GPT-4 treatment decisions were in substantial agreement with HT treatment decisions, a finding that we were able to reproduce in our experiments. Similarly, in studies on tasks beyond therapeutic decision-making, such as answering board exam questions [1,23] and diagnosing complex clinical cases [2,24,25], LLMs performed particularly well when the input data were concise and information-dense.

Basic research has indicated that LLMs struggle with lengthy texts [26] spanning over multiple prompts, potentially leading to memory loss [27] and texts with a low signal-to-noise ratio [28]. A study by Levy et al [29] demonstrated that LLM reasoning performance declined notably with increasing input length. Specifically, the authors observed a 26% drop in LLM performance when the input length was artificially increased from 250 to 3000 tokens, that is, a range of input lengths comparable to that in our study (Table S3 in Multimedia Appendix 1).

Recently, Hager et al [30] investigated the ability of LLMs to correctly diagnose patients presenting to the emergency department with abdominal pain. In this study, it was shown that deliberately withholding relevant clinical information from the LLMs paradoxically improved their diagnostic accuracy. Overall, this implies that LLMs are sensitive to both the signal-to-noise ratio and the sheer quantity of information provided.

LLMs Are Not Yet Ready for Clinical Decision-Making

The results obtained with preprocessed patient data in our study and in previous studies demonstrate the potential of LLMs in medicine. However, the use of curated and preprocessed data does not reflect the clinical situation: To this day, the communication of clinical data within hospitals is largely based on unstructured free text.

Health care professionals have high expectations of artificial intelligence (AI) to reduce their workload. This is not the case when physicians must manually extract and prepare key patient data for LLMs, as data extraction, not the actual decision-making task, is usually the most labor-intensive step.

Once key patient data have been extracted and prepared as input, simpler machine learning models (eg, tree-based models) could be used alternatively to provide decision support. In our study, as well as in the study by Salihu et al [8], simple reference models performed comparably to GPT-4, suggesting that non-LLM models could outperform LLMs if trained appropriately. In addition, nongenerative models do not exhibit undesirable behaviors, such as hallucinations and unreliability [21,31,32], and provide explainability and established measures of uncertainty quantification, which are 2 hallmarks of reasonable AI [33] that are currently not adequately implemented for LLMs [34-36].

Another hallmark of reasonable AI is to address algorithmic bias [37]. It is conceivable that the bias we observed in virtually all LLMs in our study could be due to LLMs being exposed to an abundance of TAVR-related internet literature during training compared to SAVR, subsequently influencing the treatment decisions.

A reasonable approach could be to use LLMs to extract clinical data [38] and generate input for downstream deterministic models, which then perform the decision-making. While this strategy should ideally exploit the strengths of LLMs and well-established machine learning classifiers, its effectiveness remains to be proven in future studies.

Limitations

Our study has some limitations, including a small patient cohort from a single center and the retrospective nature of our investigation. Nevertheless, the size of our study cohort (n=80) was comparable to previous key publications [2,39] studying the performance of LLMs in medicine, and we assume that our patient cohort was sufficiently large given the clear trends we observed.

The HT decisions against which we compared the LLMs' treatment decisions may themselves be nonobjective and deviate from the CPGs. We manually reviewed the HT treatment decisions and found no substantial deviations from the CPGs. Since treatment decisions are ultimately made by a team of physicians (ie, human individuals), the ground truth in experiments such as ours is inherently susceptible to some degree of subjectivity.

Given the limited cohort size and the considerable length of the medical reports, few-shot prompting or fine-tuning was not a viable option. We did not employ more sophisticated prompting techniques, such as chain-of-thought [40], and confined hyperparameter tuning to the temperature parameter. Moreover, given the rapid pace of LLM development, it is plausible that the most recently released reasoning-focused models (eg, GPT-o3 and Grok 4) may outperform those evaluated in our

study. Accordingly, our findings should be interpreted as a reflection of the current state of model performance.

The majority of LLMs evaluated were primarily trained on English-language data. While recent studies suggest that newer models exhibit greater language agnosticism, it remains plausible that our use of German-language clinical reports contributed to reduced model performance, thereby limiting the generalizability of our findings to other languages and clinical settings [41].

We acknowledge that the off-the-shelf LLMs used in our study may exhibit biases due to the underrepresentation of certain ethnic, gender, or socioeconomic groups in their training data. However, given the limited size of our cohort, we were not able to systematically assess or stratify model performance across these dimensions.

Lastly, we did not investigate whether incorporating imaging data as additional input for multimodal LLMs, such as GPT-4o, could have improved model performance in our task. While this is theoretically plausible, recent research suggests that the effectiveness of multimodal models in clinical applications depends heavily on the quality of the accompanying textual context [42,43]. Given that relevant imaging findings were generally described in detail in the imaging reports, we assume that the inclusion of imaging data in our specific use case would likely have had only a limited impact on overall model performance.

Conclusions

Our experiments are among the most challenging tasks LLMs have been asked to perform in the medical sciences. Overall, we conclude that LLMs are currently not suitable as decision makers for the treatment of patients with severe AS, as our results suggest that LLMs require elaborate preprocessing of patient data to make guideline-concordant treatment decisions. Thus, we do not share the medical community's concern that staff will be replaced by AI [44] in clinical decision-making in the near future.

Our findings suggest that LLMs should be used cautiously, particularly by medical laypersons seeking medical advice, such as second opinions. Users without extensive domain knowledge may receive treatment recommendations at a level similar to our raw experiments. This is because medical laypersons may not be able to support prompts with guideline knowledge or create case summaries of sufficient quality but will only be able to use original medical reports. The findings in the study by Hager et al [30] suggest that LLMs perform poorly when collecting additional patient data sequentially, as physicians would during a patient-physician dialogue. This suggests that the alternative to our approach—not providing all clinical data to the LLM at once, but having medical laypersons provide essential information incrementally during a chat session—is also likely to lead to suboptimal therapeutic recommendations.

Finally, medical laypersons may not be able to recognize hallucinations as effectively as medical professionals. This, combined with the eloquent and persuasive linguistic style of most LLMs, has the potential to mislead users by creating an illusion of greater certainty than warranted, aggravating the hazardous effects of incorrect treatment recommendations.

Acknowledgments

We thank Michael Gudo (MORPHISTO GmbH) for providing access to GPT-4 and Hadi El Ali (BSc), University of Bayreuth, for contributing to the illustration of Figure 1.

This work was supported by the German Centre for Cardiovascular Research (DZHK), funded by the German Federal Ministry of Education and Research, and the Charité – Universitätsmedizin Berlin. DH received 2 grants from the DZHK (grant number: 81X3100214 and grant number: 81X3100220). TR and DH are participants in the BIH Charité Digital Clinician Scientist Program funded by the Charité – Universitätsmedizin Berlin and the Berlin Institute of Health at Charité (BIH).

Data Availability

The anonymized data underlying this article will be shared upon reasonable request to the corresponding author.

Authors' Contributions

Conceptualization: TR, MH, DH, AM (equal)
Data curation: DH, FR
Formal analysis: TR, MH (equal)
Methodology: TR, MH (equal)
Supervision: AM
Visualization: MH (lead), TR (supporting)
Writing – original draft: TR, MH (equal)
Writing – review & editing: TR, MH (equal), AM (supporting), NH (supporting), TDT (supporting), MIG (supporting), AU (supporting), CK (supporting), JK (supporting), HD (supporting), BOB (supporting), GH (supporting), FB (supporting), VF (supporting)

Conflicts of Interest

DH reports financial engagements beyond the scope of the presented work. These activities include consultation services and speaking engagements for companies, including AstraZeneca, Bayer Vital, Boehringer Ingelheim, Coliquio, and Novartis.
TDT holds shares of Microsoft, Amazon, and Palantir Technologies.
AU serves as a physician proctor to Boston Scientific, Edwards Lifesciences, and Medtronic.
JK reports personal fees from Edwards and personal fees from LSI outside the submitted work.
BOB declares research funding from the British Heart Foundation and the National Institute for Health Science Research, and relevant financial activities outside the submitted work with Teleflex and Abiomed in relation to consultancy fees.
FB reports funding from Medtronic and grants from the German Federal Ministry of Education and Research, grants from the German Federal Ministry of Health, grants from the Berlin Institute of Health, personal fees from Elsevier Publishing, grants from Hans Böckler Foundation, other funds from Robert Koch Institute, grants from Einstein Foundation, and grants from Berlin University Alliance outside the submitted work.
VF declares relevant financial activities outside the submitted work with Medtronic GmbH, Biotronik SE & Co, Abbott GmbH & Co KG, Boston Scientific, Edwards Lifesciences, Berlin Heart, Novartis Pharma GmbH, JOTEC GmbH, and Zurich Heart in relation to educational grants (including travel support), fees for lectures and speeches, fees for professional consultation, and research and study funds.
AM declares receiving consulting and lecturing fees from Medtronic, lecturing fees from Bayer, and consulting fees from Pfizer. AM is the founder and managing director of x-cardiac GmbH.
The other authors have no conflicts of interest to disclose.

Multimedia Appendix 1
Additional information to support the study findings.
[DOCX File, 385 KB - xmed_v6i1e74899_app1.docx]

References

1. Kung TH, Cheatham M, Medenilla A, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLOS Digit Health 2023 Feb;2(2):e0000198. [doi: 10.1371/journal.pdig.0000198] [Medline: 36812645]
2. Kanjee Z, Crowe B, Rodman A. Accuracy of a generative artificial intelligence model in a complex diagnostic challenge. JAMA 2023 Jul 3;330(1):78-80. [doi: 10.1001/jama.2023.8288] [Medline: 37318797]
3. Tu T, Palepu A, Schaekermann M, et al. Towards conversational diagnostic AI. arXiv. Preprint posted online on Jan 11, 2024 URL: https://arxiv.org/abs/2401.05654 [accessed 2025-10-01]
4. Sorin V, Klang E, Sklair-Levy M, et al. Large language model (ChatGPT) as a support tool for breast tumor board. NPJ Breast Cancer 2023 May 30;9(1):44. [doi: 10.1038/s41523-023-00557-8] [Medline: 37253791]
5. Aghamaliyev U, Karimbayli J, Giessen-Jung C, et al. ChatGPT's gastrointestinal tumor board tango: a limping dance partner? Eur J Cancer 2024 Jul;205:114100. [doi: 10.1016/j.ejca.2024.114100] [Medline: 38729055]
6. Kozel G, Gurses ME, Gecici NN, et al. Chat-GPT on brain tumors: an examination of artificial intelligence/machine learning's ability to provide diagnoses and treatment plans for example neuro-oncology cases. Clin Neurol Neurosurg 2024 Apr;239(108238):108238. [doi: 10.1016/j.clineuro.2024.108238] [Medline: 38507989]
7. Lukac S, Dayan D, Fink V, et al. Evaluating ChatGPT as an adjunct for the multidisciplinary tumor board decision-making in primary breast cancer cases. Arch Gynecol Obstet 2023 Dec;308(6):1831-1844. [doi: 10.1007/s00404-023-07130-5] [Medline: 37458761]
8. Salihu A, Meier D, Noirclerc N, et al. A study of ChatGPT in facilitating heart team decisions on severe aortic stenosis. EuroIntervention 2024 Apr 15;20(8):e496-e503. [doi: 10.4244/EIJ-D-23-00643] [Medline: 38629422]
9. Roth GA, Mensah GA, Johnson CO, et al. Global burden of cardiovascular diseases and risk factors, 1990-2019: update from the GBD 2019 study. J Am Coll Cardiol 2020 Dec 22;76(25):2982-3021. [doi: 10.1016/j.jacc.2020.11.010] [Medline: 33309175]
10. Vahanian A, Beyersdorf F, Praz F, et al. 2021 ESC/EACTS Guidelines for the management of valvular heart disease. Eur Heart J 2022 Feb 12;43(7):561-632. [doi: 10.1093/eurheartj/ehab395] [Medline: 34453165]
11. Ye J, Chen X, Xu N, et al. A comprehensive capability analysis of GPT-3 and GPT-3.5 series models. arXiv. Preprint posted online on Dec 23, 2023 URL: https://arxiv.org/abs/2303.10420 [accessed 2025-10-01]
12. Achiam J, Adler S, Agarwal S, et al. GPT-4 technical report. arXiv. Preprint posted online on Mar 4, 2024 URL: https://arxiv.org/abs/2303.08774 [accessed 2025-10-01]
13. Anil R, Dai AM, Firat O, et al. PaLM 2 technical report. arXiv. Preprint posted online on Sep 13, 2023 URL: https://arxiv.org/abs/2305.10403 [accessed 2025-10-01]
14. Guo D, Yang D, Zhang H, et al. DeepSeek-R1: incentivizing reasoning capability in LLMs via reinforcement learning. arXiv. Preprint posted online on Jan 22, 2025 URL: https://arxiv.org/abs/2501.12948 [accessed 2025-10-01]

15. Jiang AQ, Sablayrolles A, Mensch A, et al. Mistral 7B. arXiv. Preprint posted online on Oct 10, 2023 URL: https://arxiv.org/abs/2310.06825 [accessed 2025-10-01]
16. Touvron H, Martin L, Stone K, et al. Llama 2: open foundation and fine-tuned chat models. arXiv. Preprint posted online on Jul 19, 2023 URL: https://arxiv.org/abs/2307.09288 [accessed 2025-10-01]
17. Luo R, Sun L, Xia Y, et al. BioGPT: generative pre-trained transformer for biomedical text generation and mining. Brief Bioinformatics 2022 Nov 19;23(6):bbac409. [doi: 10.1093/bib/bbac409]
18. Liévin V, Hother CE, Motzfeldt AG, Winther O. Can large language models reason about medical questions? arXiv. Preprint posted online on Dec 24, 2022 URL: https://arxiv.org/abs/2207.08143 [accessed 2025-10-01]
19. Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. J Chiropr Med 2016 Jun;15(2):155-163. [doi: 10.1016/j.jcm.2016.02.012] [Medline: 27330520]
20. Virtanen MPO, Eskola M, Jalava MP, et al. Comparison of outcomes after transcatheter aortic valve replacement vs surgical aortic valve replacement among patients with aortic stenosis at low operative risk. JAMA Netw Open 2019 Jun 5;2(6):e195742. [doi: 10.1001/jamanetworkopen.2019.5742] [Medline: 31199448]
21. Huang L, Yu W, Ma W, et al. A survey on hallucination in large language models: principles, taxonomy, challenges, and open questions. arXiv. Preprint posted online on Nov 19, 2024 URL: https://arxiv.org/abs/2311.05232 [accessed 2025-10-01]
22. Cai Y, Wang L, Wang Y, et al. MedBench: a large-scale chinese benchmark for evaluating medical large language models. arXiv. Preprint posted online on Dec 20, 2023 URL: https://arxiv.org/abs/2312.12806 [accessed 2025-10-01]
23. Gilson A, Safranek CW, Huang T, et al. How does ChatGPT perform on the United States Medical Licensing Examination (USMLE)? The implications of large language models for medical education and knowledge assessment. JMIR Med Educ 2023 Feb 8;9:e45312. [doi: 10.2196/45312] [Medline: 36753318]
24. Eriksen AV, Möller S, Ryg J. Use of GPT-4 to diagnose complex clinical cases. NEJM AI 2024 Jan;1(1):AIp2300031. [doi: 10.1056/AIp2300031]
25. Novak A, Zeljković I, Rode F, et al. The pulse of artificial intelligence in cardiology: a comprehensive evaluation of state-of-the-art large language models for potential use in clinical cardiology. medRxiv. Preprint posted online on Dec 7, 2024 URL: https://www.medrxiv.org/content/10.1101/2023.08.08.23293689v3 [accessed 2025-10-01] [doi: 10.1101/2023.08.08.23293689]
26. Liu NF, Lin K, Hewitt J, et al. Lost in the middle: how language models use long contexts. arXiv. Preprint posted online on Nov 20, 2023 URL: https://arxiv.org/abs/2307.03172 [accessed 2025-10-01]
27. Sejnowski TJ. Large language models and the reverse turing test. Neural Comput 2023 Feb 17;35(3):309-342. [doi: 10.1162/neco_a_01563] [Medline: 36746144]
28. Wang B, Wei C, Liu Z, Lin G, Chen NF. Resilience of large language models for noisy instructions. arXiv. Preprint posted online on Oct 3, 2024 URL: https://arxiv.org/abs/2404.09754 [accessed 2025-10-01]
29. Levy M, Jacoby A, Goldberg Y. Same task, more tokens: the impact of input length on the reasoning performance of large language models. arXiv. Preprint posted online on Jul 10, 2024 URL: https://arxiv.org/abs/2402.14848 [accessed 2025-10-01]
30. Hager P, Jungmann F, Holland R, et al. Evaluation and mitigation of the limitations of large language models in clinical decision-making. Nat Med 2024 Sep;30(9):2613-2622. [doi: 10.1038/s41591-024-03097-1] [Medline: 38965432]
31. Singhal K, Azizi S, Tu T, et al. Large language models encode clinical knowledge. Nature New Biol 2023 Aug;620(7972):172-180. [doi: 10.1038/s41586-023-06291-2] [Medline: 37438534]
32. Roustan D, Bastardot F. The clinicians' guide to large language models: a general perspective with a focus on hallucinations. Interact J Med Res 2025 Jan 28;14:e59823. [doi: 10.2196/59823] [Medline: 39874574]
33. Sivarajah U, Wang Y, Olya H, Mathew S. Responsible artificial intelligence (AI) for digital health and medical analytics. Inf Syst Front 2023 Jun 5;5(1–6):1-6. [doi: 10.1007/s10796-023-10412-7] [Medline: 37361886]
34. Luo H, Specia L. From understanding to utilization: a survey on explainability for large language models. arXiv. Preprint posted online on Feb 22, 2024 URL: https://arxiv.org/abs/2401.12874 [accessed 2025-10-01]
35. Liu L, Pan Y, Li X, Chen G. Uncertainty estimation and quantification for LLMs: a simple supervised approach. arXiv. Preprint posted online on Oct 23, 2024 URL: https://arxiv.org/abs/2404.15993 [accessed 2025-10-01]
36. Quttainah M, Mishra V, Madakam S, Lurie Y, Mark S. Cost, usability, credibility, fairness, accountability, transparency, and explainability framework for safe and effective large language models in medical education: narrative review and qualitative study. JMIR AI 2024 Apr 23;3:e51834. [doi: 10.2196/51834] [Medline: 38875562]
37. Kim J, Vajravelu BN. Assessing the current limitations of large language models in advancing health care education. JMIR Form Res 2025 Jan 16;9:e51319. [doi: 10.2196/51319] [Medline: 39819585]
38. Dagdelen J, Dunn A, Lee S, et al. Structured information extraction from scientific text with large language models. Nat Commun 2024 Feb 15;15(1):1418. [doi: 10.1038/s41467-024-45563-x] [Medline: 38360817]
39. Huang AS, Hirabayashi K, Barna L, Parikh D, Pasquale LR. Assessment of a large language model's responses to questions and cases about glaucoma and retina management. JAMA Ophthalmol 2024 Apr 1;142(4):371-375. [doi: 10.1001/jamaophthalmol.2023.6917] [Medline: 38386351]
40. Wei J, Wang X, Schuurmans D, et al. Chain-of-thought prompting elicits reasoning in large language models. Presented at: NIPS '22: Proceedings of the 36th International Conference on Neural Information Processing Systems; Nov 28 to Dec 9, 2022; New Orleans, LA, USA. [doi: 10.5555/3600270.3602070]

41. Rosoł M, Gąsior JS, Łaba J, Korzeniewski K, Młyńczak M. Evaluation of the performance of GPT-3.5 and GPT-4 on the Polish Medical Final Examination. Sci Rep 2023 Nov 22;13(1):20512. [doi: 10.1038/s41598-023-46995-z] [Medline: 37993519]
42. Günay S, Öztürk A, Özerol H, Yiğit Y, Erenler AK. Comparison of emergency medicine specialist, cardiologist, and chat-GPT in electrocardiography assessment. Am J Emerg Med 2024 Jun;80:51-60. [doi: 10.1016/j.ajem.2024.03.017] [Medline: 38507847]
43. Zeljkovic I, Novak A, Lisicic A, et al. Beyond text: the impact of clinical context on GPT-4's 12-lead electrocardiogram interpretation accuracy. Can J Cardiol 2025 Jul;41(7):1406-1414. [doi: 10.1016/j.cjca.2025.01.036] [Medline: 39971004]
44. Fogo AB, Kronbichler A, Bajema IM. AI's threat to the medical profession. JAMA 2024 Feb 13;331(6):471-472. [doi: 10.1001/jama.2024.0018] [Medline: 38241042]

Abbreviations

AI: artificial intelligence
AS: aortic stenosis
CPG: clinical practice guideline
CT: computed tomography
EACTS: European Association for Cardio-Thoracic Surgery
ESC: European Society of Cardiology
FBI: frequency bias index
HT: heart team
ICC: intraclass correlation coefficient
LLM: large language model
SAVR: surgical aortic valve replacement
TAVR: transcatheter aortic valve replacement

Edited by A Grover; submitted 26.03.25; peer-reviewed by A Novak, R Singh; revised version received 23.07.25; accepted 19.08.25; published 03.11.25.

Please cite as:
Roeschl T, Hoffmann M, Hashemi D, Rarreck F, Hinrichs N, Trippel TD, Gröschel MI, Unbehaun A, Klein C, Kempfert J, Dreger H, O'Brien B, Hindricks G, Balzer F, Falk V, Meyer A
Assessing the Limitations of Large Language Models in Clinical Practice Guideline–Concordant Treatment Decision-Making on Real-World Data: Retrospective Study
JMIRx Med 2025;6:e74899
URL: https://xmed.jmir.org/2025/1/e74899
doi:10.2196/74899

Willingness to Pay for the COVID-19 Vaccine and Its Correlates in Bangladesh: Cross-Sectional Study

Mohammad Bellal Hossain[1], PhD; Md Zakiul Alam[1], MSS; Md Syful Islam[1], MSS; Shafayat Sultan[1], MSS; Md Mahir Faysal[1], MSS; Sharmin Rima[1], MSS; Md Anwer Hossain[2], MSS; Abdullah Al Mamun[3], MSS; Abdullah-Al- Mamun[4], PhD

[1]Department of Population Sciences, University of Dhaka, Third Floor, Arts Faculty Building, Dhaka, Bangladesh

[2]Laboratory of Fertility and Wellbeing, Max Planck Institute for Demographic Research, Rostock, Germany

[3]Department of Social Relations, East West University, Dhaka, Bangladesh

[4]Department of Japanese Studies, University of Dhaka, Dhaka, Bangladesh

Corresponding Author:
Mohammad Bellal Hossain, PhD
Department of Population Sciences, University of Dhaka, Third Floor, Arts Faculty Building, Dhaka, Bangladesh

Related Articles:

Companion article: https://www.medrxiv.org/content/10.1101/2024.12.03.24318442v1

Companion article: https://med.jmirx.org/2025/1/e79353

Companion article: https://med.jmirx.org/2025/1/e79354

Companion article: https://med.jmirx.org/2025/1/e79355

Companion article: https://med.jmirx.org/2025/1/e79352

Abstract

Background: The Government of Bangladesh offers COVID-19 vaccines at no cost; however, sustaining this free vaccination program for a large population poses significant challenges. Thus, assessing the willingness to pay (WTP) for the COVID-19 vaccine is essential for understanding potential pricing strategies, subsidy requirements, and vaccine demand.

Objective: This study aimed to assess the prevalence of WTP for the COVID-19 vaccine and its correlates.

Methods: A cross-sectional design was used to collect data from 1497 respondents through web-based platform and face-to-face interviews. Multivariable logistic regression was used to analyze the correlates of the WTP.

Results: The results showed that 772 of 1497 (51.6%) participants were willing to pay for the COVID-19 vaccine, with a median of 300 BDT (IQR 150-500 BDT; a currency exchange rate of 1 BDT=US $0.008 is applicable). The WTP was significantly higher among individuals with a graduate degree (adjusted odds ratio [aOR] 1.98, 95% CI 1.14-3.45) or master's and MPhil or PhD-level education (aOR 1.93, 95% CI 1.07-3.48) and those with higher knowledge about the vaccine (aOR 1.09, 95% CI 1.02-1.15), positive behavioral practices (aOR 1.11, 95% CI 1.06-1.17), stronger subjective norms regarding COVID-19 vaccine (aOR 1.25, 95% CI 1.08-1.46), and higher anticipated regret of getting infected with COVID-19 (aOR 1.17, 95% CI 1.04-1.32). Conversely, WTP was lower among participants with negative attitudes toward vaccines (aOR 0.91, 95% CI 0.88 - 0.95) and high perceived behavioral control regarding COVID-19 vaccination (aOR 0.86, 95% CI 0.76 - 0.96; *P*=.006).

Conclusions: With nearly half of the respondents unwilling to pay, this study highlights the need to improve vaccine-related knowledge and enhance income-based affordability to increase WTP. Health promotion efforts should focus on disseminating knowledge about vaccines and addressing negative perceptions. Additionally, a subsidized program for low-income groups can help mitigate financial barriers and promote equitable access to vaccines.

(JMIRx Med 2025;6:e69827) doi:10.2196/69827

KEYWORDS

Bangladesh; willingness to pay; vaccines; COVID-19; infectious diseases; public health; public safety; cross-sectional study; financial

Introduction

COVID-19 has caused tremendous health and socioeconomic challenges worldwide. As the treatment of COVID-19 is relatively expensive, preventing COVID-19 with a safe and effective vaccine is of utmost importance [1], which is considered one of the most successful public health interventions [2]. Even with safe and effective vaccines, the success of vaccination coverage programs relies on several other factors, such as willingness to vaccinate [3-6]. The intention to be vaccinated and the success of vaccination coverage programs have been influenced by the economic considerations of many people [7-9]. In this regard, willingness to pay (WTP) has emerged as a concept defined as the WTP any amount of money for a vaccine, health services, or technology [10-12].

To combat the vulnerability caused by COVID-19, the Government of Bangladesh (GoB) launched its largest vaccination program, providing vaccines free of charge. However, sustaining this free vaccination program is challenging [13] for a resource-poor country with a large population, such as Bangladesh. On the other hand, from the demand-side perspective, individuals' out-of-pocket health expenditures in Bangladesh are already 74%, and 18.7% of the total population lives below the poverty line [14,15], making full payment for vaccines impossible. Thus, assessing WTP can offer insights into public demand and guide GoBs' future pricing and payment strategies [1,3].

Few studies in Bangladesh have measured WTP in non–COVID-19 situations [8,16,17], whereas there is considerable evidence assessing WTP in the COVID-19 context in various countries [7,11,18]. WTP, both in COVID-19 and non–COVID-19 contexts, is determined by socioeconomic and demographic factors [3,8,11,19,20], health-related factors, knowledge about the disease and vaccine [7], different constructs of the health belief model, and behavioral factors [12].

However, there is scant evidence regarding WTP for the COVID-19 vaccine in Bangladesh. Existing studies conducted in Bangladesh to assess the prevalence and median WTP toward the COVID-19 vaccine [21] and associated factors [22] lacked representativeness of the Bangladeshi population, as the studies were only web-based and may result in an overrepresentation of more urban, younger, or technology-savvy individuals, skewing the sample and making it unreflective of the entire population. Thus, this study aimed to assess the prevalence of WTP for the COVID-19 vaccine and its correlates.

Methods

Study Design and Data Collection

This study used a cross-sectional design. The calculated sample size was 1635 using the formula (z^2pq/e^2)*Deff*NR. The *z* score for a 95% CI was 1.96, and the prevalence of willingness to accept a COVID-19 vaccine in an earlier study was 32.5% (*p*) [23]. We used a margin of error, *e*=0.03, for the sampling variation design effect (Deff)=1.6 and a 10% nonresponse rate. The ratio for face-to-face and web-based platform surveys was 2:1, considering the country's digital divide. The overall participation rate was 93.1% (face-to-face survey: 91.9% and web-based platform survey: 97.7%), and 112 respondents did not consent to participate in the study (web-based platform survey: n=11 and face-to-face survey: n=101). Finally, 26 respondents who were unaware of the COVID-19 vaccine were excluded from the sample. Thus, the final sample for this study was 1497 for analysis (web-based platform survey: n=475, 31.7% and face-to-face survey: n=1022, 68.3%). As the GoB intended to initiate a mass vaccination program from February 7, 2021, we fixed February 1, 2021, to February 7, 2021, for data collection time.

Data were collected using a web-based platform through face-to-face interviews through Google Forms, where the questionnaire was developed in Bengali. Google Forms collects web-based data by prioritizing user friendliness and accessible design. This study conducted a pretest to validate and further improve the survey questionnaire. During the pretest, we tested the technical functionality and usability of the questionnaire. This helped us to identify and address missing questions, response codes, chronology, navigation, and technical performance. Due to the government's restrictions on movement, safety protocols, and public hesitancy to meet data collectors at their homes, we decided to collect web-based data first to quickly reach a large audience. The survey link was circulated among the research team members' networks via email, Facebook, WhatsApp, and other platforms. The respondents were also asked to share the link with their networks to reach as many people as possible. We used a single-response-per-participant option for the web-based survey to prevent duplicate entries. The web-based questionnaire contained 12 sections, each displayed on a separate screen for clarity. Upon completion, the respondents received a confirmation message indicating the successful submission of their responses. Additionally, the questionnaire included the respondents' options to request the study results. Data collection for this study was completed between February 1 and 7, 2021 (web-based data were collected on February 1 and 3, 2021). After 3 days, the data were checked for divisional and age- and sex-specific distributions. To ensure representativeness of the sample, face-to-face interviews were conducted to determine the population's national representation in terms of age, sex, residence, division, and marital status. Data were collected from 2 randomly selected districts in each of 8 divisions. Four days were allocated to collecting data through face-to-face interviews, and the interviewers maintained proper health measures during the interviews.

For offline data collection, the selection criteria to participate in this study were at least 18 years and older of age and knew about the COVID-19 vaccine, with an additional criterion for the web-based survey respondents' reading and writing ability.

Measures

Outcome Variables

Two questions were used to assess respondents' WTP. The first question was, "Would you like to pay for the COVID-19 vaccine?" with a binary response (yes or no). If the response to the first question was "yes," then the second question was,

"What is the maximum amount you are willing to pay for the COVID-19 vaccine?".

Independent Variables

Selection Process of Independent Variables

The independent variables in this study were selected through an extensive review of the literature. A range of socioeconomic, demographic, health status, COVID-19 infection, knowledge about the COVID-19 vaccine and vaccination process, conspiracy theories, preventive behavioral practices, and health behavioral models were selected. In health-related research, the theory of planned behavior (TPB), the health belief model, and 5C psychological antecedents are frequently used to capture attitudinal and normative influences on health behavior; however, we selected only the TPB because of its better predictability in such instances [24].

Socioeconomic and Demographic Variables

Several socioeconomic and demographic variables such as age, sex, religion, marital status, place of residence, household income, and occupation were used as independent variables in this study.

Perceived Health Status and COVID-19 Infection

Respondents' perceived health status was assessed using the question, "How would you like to rate your health?" The response options were categorized as 1=bad, 2=moderate, and 3=good. Respondents were asked 3 separate binary (yes or no) questions regarding their experience with infection to assess their COVID-19 infection status: "Have you ever been infected with COVID-19? Has any member of your family been infected with COVID-19? Has any friend or acquaintance of yours contracted COVID-19?"

Knowledge of the COVID-19 Vaccine

Knowledge of the COVID-19 vaccine was assessed using four 5-point Likert-type items. The scores range from 1 to 20, with higher scores indicating greater knowledge. Cronbach α was 0.643. The commonly accepted Cronbach α threshold for measuring the internal consistency of the scales was 0.70. However, previous literature suggests that acceptance of moderate internal consistency is acceptable when measuring perceptual or cognitive-related issues, such as knowledge or susceptibility, and Cronbach α values between 0.60 and 0.70 can still be considered sufficient for complex psychological and health-related measures [25].

Knowledge About the Vaccination Process

Six questions with binary (yes=1 and no=0) response options were used to measure knowledge of the COVID-19 vaccination process, with a total score ranging from 0 to 6. Cronbach α was 0.765, indicating good internal consistency among the 6 questions. The higher the score, the better the understanding of the vaccination process.

COVID-19 Vaccine Conspiracy

Nine 5-point Likert-type items were used to assess conspiracy beliefs related to the COVID-19 vaccine [23], yielding scores ranging from 9 to 45, where a higher score indicated a stronger belief in conspiracy theories regarding the COVID-19 vaccine. Reliability analysis of these 9 questions revealed good internal consistency (α=.716).

Preventive Behavioral Practices Related to COVID-19

Preventive behavioral practices related to COVID-19 were measured using three 4-point items, with a composite value ranging from 1 to 12. A higher score indicates better preventive practices against COVID-19, and the items demonstrated excellent internal reliability (α=.857).

Theory of Planned Behavior

The TPB comprises 4 domains: attitude toward the COVID-19 vaccine, subjective norms toward the COVID-19 vaccine, perceived behavioral control against COVID-19 vaccination, and anticipated regret regarding getting infected by COVID-19.

Attitude Toward COVID-19 Vaccine

A 6-item 5-point Likert-type scale was used to assess attitudes related to the COVID-19 vaccine, with scores ranging from 6 to 30. A higher score on this scale indicates a more negative attitude toward the COVID-19 vaccine. Reliability analysis of these 6 items revealed good internal consistency (α=.739).

Subjective Norms Toward the COVID-19 Vaccine

Subjective norm refers to an individual's perception of social pressure to perform or not perform a behavior. The subjective norm regarding the COVID-19 vaccine was assessed using a single 5-point item: "I believe my family members will support me in getting vaccinated against COVID-19."

Perceived Behavioral Control Against COVID-19 Vaccination

Perceived behavioral control is defined as an individual's belief about the ease or difficulty of performing a particular behavior, which can reflect experience and anticipated obstacles. It was measured using a single 5-point item: "I can register for COVID-19 vaccination if I want."

Anticipated Regret Regarding Getting Infected by COVID-19

The anticipated regret of not getting vaccinated was assessed using a single 5-point item: "If I do not get a COVID-19 vaccine and end up getting Coronavirus, I will regret not getting the vaccination."

Statistical Analysis

At the beginning of the data analysis, we used poststratification weighting techniques to reduce biases stemming from sampling, thereby making our study representative of the national population of Bangladesh. We used a weight-adjustment technique for the age variable using the following formula:

$$\omega i = pisi$$

where ωi is the weight-adjusted factor, pi is the relative proportion of population characteristics, and si is the proportion of sample characteristics. For example, the sample proportion for individuals aged 18‐24 years was 0.289, whereas the national population proportion, according to Bangladesh's 2022 Population and Housing Census, was 0.201. Thus, using the formula, we calculated a weight of 1.43 for the age group 18‐24 years. Similarly, we calculated the weight of each age group to ensure that our study population was representative of

the national population in terms of age distribution. The weighted samples were used for further analysis.

We first used univariate descriptive statistics (percentage, mean, and SD) for all variables to obtain the background characteristics of the participants and the prevalence of WTP for the COVID-19 vaccine. At the bivariate level, we obtained differentials in WTP for the COVID-19 vaccine using a chi-square test and point biserial correlation. Statistically significant variables ($P \leq .05$) at the bivariate level were entered into a hierarchical logistic regression model to assess the correlates of WTP for the COVID-19 vaccine. We used only descriptive analysis (mean and median) for WTP the highest amount of money. Data were analyzed using SPSS (version 26; IBM Corp).

Ethical Considerations

Ethics approval was obtained from the Bangladesh Medical Research Council (registration 39131012021). Voluntary participation was encouraged, and no incentives were provided to the participants. First, the aims, objectives, potential scopes, and implications of the study's findings were communicated to the participants, and they then provided written consent. The respondents then participated in the study. Privacy and confidentiality were maintained throughout the study by ensuring that all participant data were anonymized and securely stored. Only authorized research personnel had access to the data, and the participants' identities were kept confidential.

Results

Background Characteristics of the Participants

The background characteristics of the respondents are presented in Table 1, which shows that 340 of 1497 (22.7%) respondents in this study were between 31 and 39 years. Of the 1497 participants, 808 (54%) were male. In terms of religion, 1297 (86.7%) participants were Muslim, and 1061 (71%) participants were married. In total, 1129 (75.3%) had at least a secondary education and higher education. A total of 979 (66.4%) were from rural areas, and the highest percentage of respondents (n=474, 31.7%) were from the Dhaka division (Table 1). In total, 395 (26.4%) were homemakers. The mean number of household members was 4.95, and they had a mean collective monthly family income of 38,500 BDT (a currency exchange rate of 1 BDT=US $0.008 is applicable). Most of the participants (n=1074, 71.8%) identified themselves as having perfect health status (Table 1).

Table . Unweighted and weighted background characteristics of the respondents (N=1497).

Variables		Unweighted study sample	Weighted study sample
Age (years), n (%)			
	18 - 24	432 (28.9)	300 (20.1)
	25 - 30	362 (24.2)	295 (19.7)
	31 - 39	254 (17.0)	340 (22.7)
	40 - 49	236 (15.8)	276 (18.5)
	50+	213 (14.2)	286 (19.1)
Sex, n (%)			
	Female	692 (46.2)	689 (46.0)
	Male	805 (53.8)	808 (54.0)
Religion, n (%)			
	Muslim	1301 (86.9)	1297 (86.7)
	Others	196 (13.1)	200 (13.3)
Marital status, n (%)			
	Unmarried	575 (38.4)	436 (29.1)
	Married	922 (61.6)	1061 (70.9)
Education, n (%)			
	No education	129 (8.6)	158 (10.6)
	Primary	179 (12.0)	210 (14.1)
	Secondary and higher secondary	448 (29.9)	440 (29.4)
	Graduate	400 (26.7)	333 (22.2)
	Master's and MPhil or PhD	341 (22.8)	356 (23.7)
Place of residence, n (%)			
	Rural	963 (64.3)	979 (66.4)
	Urban (other than city corporation)	179 (12.0)	171 (11.4)
	City corporation	355 (23.7)	347 (23.2)
Administrative division, n (%)			
	Barisal	114 (7.6)	118 (7.9)
	Chattogram	253 (16.9)	267 (17.8)
	Dhaka	478 (31.9)	474 (31.7)
	Khulna	137 (9.2)	132 (8.8)
	Mymensingh	108 (7.2)	106 (7.1)
	Rajshahi	180 (12.0)	179 (12.0)
	Rangpur	114 (7.6)	108 (7.2)
	Sylhet	113 (7.5)	113 (7.5)
Occupation, n (%)			
	Government, private, and NGO[a] sector job	202 (13.5)	217 (14.5)
	Professionals	277 (18.5)	311 (20.8)
	Homemakers	348 (23.2)	395 (26.4)
	Students and unemployed	473 (31.6)	352 (23.5)
	Agriculture and day laborer	102 (6.81)	121 (8.1)
	Others	95 (6.34)	101 (6.7)

Variables		Unweighted study sample	Weighted study sample
Collective monthly family income (BDT)[b], n (%)			
	Less than 10,000	261 (17.4)	267 (17.9)
	10,000 - 20,000	394 (26.3)	390 (26.1)
	20,000 - 30,000	271 (18.1)	265 (17.7)
	30,000 - 40,000	161 (10.8)	158 (10.6)
	40,000 and above	410 (27.4)	417 (27.8)
Perceived health status, n (%)			
	Bad or very bad	51 (3.4)	60 (4.0)
	Moderate	333 (22.2)	363 (24.3)
	Good or very good	1113 (74.3)	1074 (71.8)
Was infected with the coronavirus, n (%)			
	Yes	86 (5.7)	85 (5.7)
	No	1411 (94.3)	1412 (94.3)
Respondent’s family members were infected with the coronavirus, n (%)			
	Yes	152 (10.2)	148 (9.9)
	No	1345 (89.8)	1349 (90.1)
Respondent’s friends got infected with the coronavirus, n (%)			
	Yes	600 (40.1)	592 (39.6)
	No	897 (59.9)	905 (60.4)
Household size, mean (SD)		4.95 (1.98)	4.95 (2.05)

[a]NGO: nongovernmental organization.

[b]A currency exchange rate of 1 BDT=US $0.008 is applicable.

Prevalence of WTP for the COVID-19 Vaccine and Its Differentials

The prevalence of WTP for the COVID-19 vaccine was 772 (51.6%) participants. However, 725 (48.4%) participants refused to pay any money for the COVID-19 vaccine.

Table 2 shows that WTP varied significantly ($P \leq .05$) by religion, marital status, education, place of residence, administrative division, occupation, household income, and coronavirus infection status of respondents, family, and friends. For instance, respondents who identified as others than Muslim, unmarried, had master's and MPhil or PhD-level education, resided in an urban area, were from the Sylhet division, employed in a professional job, and had an income of 40,000 BDT and above had higher WTP at the bivariate level analysis (Table 2).

Table . Differentials of willingness to pay for COVID-19 vaccine among the study population using the chi-square test.

Variables		Willingness to pay		*P* value
		No, n (%)	Yes, n (%)	
Total		725 (48.4)	772 (51.6)	
Age (years)				.16
	18 - 24	140 (46.7)	160 (53.3)	
	25 - 30	140 (47.5)	155 (52.5)	
	31 - 39	154 (45.3)	186 (54.7)	
	40 - 49	133 (48.2)	143 (51.8)	
	50+	157 (54.9)	129 (45.1)	
Sex				.72
	Female	330 (47.9)	359 (52.1)	
	Male	394 (48.8)	414 (51.2)	
Religion				<.001
	Muslim	653 (50.3)	644 (49.7)	
	Others	72 (36.0)	128 (64.0)	
Marital status				.01
	Unmarried	188 (43.2)	247 (56.8)	
	Married	537 (50.6)	525 (49.4)	
Education				<.001
	No education	111 (69.8)	48 (30.2)	
	Primary	132 (62.9)	78 (37.1)	
	Secondary and higher secondary	227 (51.6)	213 (48.4)	
	Graduate	129 (38.7)	204 (61.3)	
	Master's and MPhil or PhD	126 (35.5)	229 (64.5)	
Place of residence				<.001
	Rural	526 (53.7)	453 (46.3)	
	Urban (other than city corporation)	61 (35.7)	110 (64.3)	
	City corporation	138 (39.8)	209 (60.2)	
Administrative division				.001
	Barisal	73 (61.9)	45 (38.1)	
	Chattogram	123 (46.1)	144 (53.9)	
	Dhaka	206 (43.5)	268 (56.5)	
	Khulna	70 (53.0)	62 (47.0)	
	Mymensingh	64 (60.4)	42 (39.6)	
	Rajshahi	87 (48.9)	92 (51.1)	
	Rangpur	55 (50.9)	53 (49.1)	
	Sylhet	47 (41.6)	66 (58.4)	
Occupation				<.001
	Government, private, and NGO[a] sector job	88 (40.4)	129 (59.6)	
	Professionals	123 (39.5)	188 (60.5)	
	Homemakers	221 (55.9)	174 (44.1)	

Variables		Willingness to pay		*P* value
		No, n (%)	Yes, n (%)	
	Students and unemployed	157 (44.7)	195 (55.3)	
	Agriculture and day laborer	77 (63.6)	44 (36.4)	
	Others	59 (58.4)	42 (41.6)	
Collective monthly family income (BDT)[b]				<.001
	Less than 10,000	156 (58.4)	111 (41.6)	
	10,000 - 20,000	232 (59.5)	158 (40.5)	
	20,000 - 30,000	127 (47.9)	138 (52.1)	
	30,000 - 40,000	59 (37.3)	99 (62.7)	
	40,000 and above	151 (36.2)	266 (63.8)	
Perceived health status				.05
	Bad or very bad	30 (50.0)	30 (50.0)	
	Moderate	195 (53.7)	168 (46.3)	
	Good or very good	499 (46.5)	575 (53.5)	
Was infected with the coronavirus				.007
	Yes	29 (34.1)	56 (65.9)	
	No	696 (49.3)	716 (50.7)	
Respondent's family members were infected with the coronavirus				.005
	Yes	55 (34.2)	93 (62.8)	
	No	669 (49.6)	680 (50.4)	
Respondent's friends got infected with the coronavirus				<.001
	Yes	233 (39.4)	359 (60.6)	
	No	492 (54.4)	413 (45.6)	

[a]NGO: nongovernmental organization.

[b]A currency exchange rate of 1 BDT=US $0.008 is applicable.

Table 3 presents the point biserial correlation among WTP and independent variables. Knowledge about the COVID-19 vaccine and vaccination process, preventive behavioral practices related to COVID-19, subjective norm toward the COVID-19 vaccine, and anticipated regret regarding getting infected by COVID-19 had a statistically significant positive correlation with WTP. On the other hand, COVID-19 conspiracy beliefs, attitude toward the COVID-19 vaccine, and perceived behavioral control against COVID-19 vaccination were negatively correlated with WTP.

Table . Point biserial correlation between willingness to pay and selected independent variables.

Variables	Correlation coefficient (*r*)	*P* value
Household size	−0.038	.15
Knowledge about the COVID-19 vaccine	0.082	.001
Knowledge about the vaccination process	0.203	<.001
COVID-19 vaccine conspiracy	−0.165	<.001
Behavioral practice related to COVID-19	0.255	<.001
Attitude toward the COVID-19 vaccine	−0.315	<.001
Perceived behavioral control against COVID-19 vaccination	−0.163	<.001
Anticipated regret regarding getting infected by COVID-19	0.190	<.001
Subjective norms toward the COVID-19 vaccine	0.221	<.001

Correlates of WTP for COVID-19 Vaccine

The significant variables at the bivariate level were entered into a hierarchical logistic regression model to assess the correlates of WTP for the COVID-19 vaccine. Three models were developed for this study. The first model was constructed using the socioeconomic and demographic variables. The second model was built using the variables from the first model along with knowledge of the COVID-19 vaccine and vaccination process, COVID-19 vaccine conspiracy theories, behavioral practices related to COVID-19, and health-related variables. Furthermore, the final model included all variables from the second model as well as components of the TPB. The final model showed that education, administrative division of Bangladesh, knowledge about the COVID-19 vaccine, behavioral practice related to COVID-19, attitude toward the COVID-19 vaccine, subjective norm toward COVID-19 vaccine, perceived behavioral control against COVID-19 vaccination, and anticipated regret regarding COVID-19 infection were statistically significant ($P \leq .05$) predictors of WTP (Table 4).

Table . Correlates of willingness to pay for COVID-19 vaccine in Bangladesh using multivariable logistic regression (N=1497).

Variables	Model 1[a], aOR[b] (95% CI)	Model 2[c], aOR (95% CI)	Model 3[d], aOR (95% CI)
Religion: other (Muslim as RC[e])	2.00 (1.43-2.81)[f]	1.75 (1.23-2.49)[g]	1.40 (0.97-2.01)[h]
Marital status: unmarried (married as RC)	1.08 (0.76-1.54)	1.09 (0.76-1.57)	1.23 (0.84-1.80)
Education (no education as RC)			
Primary	1.40 (0.89-2.22)	1.10 (0.68-1.78)	0.99 (0.60-1.62)
Secondary and higher secondary	2.09 (1.36-3.19)[g]	1.43 (0.92-2.24)	1.28 (0.80-2.04)
Graduate	3.36 (2.04-5.53)[f]	2.09 (1.23-3.55)[g]	1.98 (1.14-3.45)[h]
Master's and MPhil or PhD	3.18 (1.87-5.41)[f]	1.82 (1.03-3.22)[h]	1.93 (1.07-3.48)[h]
Place of residence (rural as RC)			
Urban (other than city corporation)	1.40 (0.95-2.05)	1.11 (0.75-1.67)	1.1 (0.73-1.69)
City corporation	1.02 (0.74-1.42)	0.96 (0.67-1.36)	1.17 (0.81-1.69)
Administrative division of Bangladesh (Barisal as RC)			
Chattogram	1.80 (1.13-2.87)[h]	2.34 (1.43-3.84)[g]	2.27 (1.61-4.68)[f]
Dhaka	1.75 (1.13-2.72)[h]	1.22 (1.40-3.54)[g]	2.72 (1.66-4.45)[f]
Khulna	1.47 (0.86-2.51)	2.23 (1.27-3.92)[g]	3.37 (1.84-6.17)[f]
Mymensingh	0.96 (0.54-1.67)	1.01 (0.55-1.83)	1.13 (0.59-2.16)
Rajshahi	1.90 (1.14-3.15)[h]	2.54 (1.50-4.31)[g]	3.36 (1.19-5.91)[f]
Rangpur	1.49 (0.84-2.65)	2.11 (1.16-3.84)[h]	3.22 (1.70-6.09)[f]
Sylhet	2.80 (1.60-4.89)[g]	3.46 (1.93-6.21)[f]	4.70 (2.53-8.74)[f]
Occupation (government, private, and NGO[i] sector job as RC)			
Professionals	1.15 (0.67-1.97)	1.04 (0.59-1.83)	1.18 (0.65-2.12)
Homemakers	1.56 (0.95-2.55)	1.40 (0.84-2.35)	1.50 (0.88-2.58)
Students and unemployed	1.58 (0.98-2.55)	1.70 (1.03-2.81)[h]	1.93 (1.14-3.25)[h]
Agriculture and day labor	1.17 (0.69-2.01)	1.21 (0.69-2.12)	1.34 (0.75-2.40)
Others	1.27 (0.71-2.28)	1.48 (0.80-2.72)	1.51 (0.80-2.85)
Income (BDT)[j] (less than 10,000 as RC)			
10,000 - 20,000	0.91 (0.65-1.27)	0.87 (0.61-1.22)	0.98 (0.68-1.40)
20,000 - 30,000	1.16 (0.80-1.68)	1.11 (0.76-1.64)	1.25 (0.84-1.87)
30,000 - 40,000	1.67 (1.07-2.59)[h]	1.56 (0.99-2.47)	1.48 (0.92-2.37)
40,000 and above	1.59 (1.07-2.37)*[h]	1.33 (0.88-2.02)	1.34 (0.87-2.06)
Knowledge about the COVID-19 vaccine	—[k]	1.13 (1.07-1.19)[f]	1.09 (1.02-1.15)[g]
Knowledge about the vaccination process	—	1.11 (1.04-1.19)[g]	1.06 (0.99-1.15)
Behavioral practice related to COVID-19	—	1.15 (1.09-1.20)[f]	1.11 (1.06-1.17)[f]
COVID-19 vaccine conspiracy	—	0.94 (0.91-0.97)[f]	1.01 (0.97-1.05)
Respondent got infected with coronavirus: yes (no as RC)	—	1.11 (0.66-1.87)	1.23 (0.71-2.12)

Variables	Model 1[a], aOR[b] (95% CI)	Model 2[c], aOR (95% CI)	Model 3[d], aOR (95% CI)
Respondent's family member got infected with coronavirus: yes (no as RC)	—	1.09 (0.73-1.65)	1.08 (0.70-1.66)
Respondents' friends or peers got infected with coronavirus: yes (no as RC)	—	1.15 (0.87-1.51)	1.16 (0.87-1.55)
Attitude toward the COVID-19 vaccine	—	—	0.91 (0.88-0.95)[f]
Subjective norms toward the COVID-19 vaccine	—	—	1.3 (1.12-1.51)[g]
Perceived behavioral control against COVID-19 vaccination	—	—	0.84 (0.75-0.94)[g]
Anticipated regret regarding getting infected by COVID-19	—	—	1.18 (1.06-1.32)[g]

[a]Constant=0.154; $P<.001$; −2 log likelihood=1924; Cox and Snell $R^2=0.095$; Nagelkerke $R^2=0.127$.

[b]aOR: adjusted odds ratio.

[c]Constant=0.028; $P<.001$; −2 log likelihood=1826; Cox and Snell $R^2=0.152$; Nagelkerke $R^2=0.203$.

[d]Constant=0.022; $P<.001$; −2 log likelihood=1730; Cox and Snell $R^2=0.206$; Nagelkerke $R^2=0.274$.

[e]RC: reference category.

[f]$P<.001$.

[g]$P<.01$.

[h]$P<.05$.

[i]NGO: nongovernmental organization.

[j]A currency exchange rate of 1 BDT=US $0.008 is applicable.

[k]Not available.

According to the final model, respondents of other religions were 40% more likely to pay for the COVID-19 vaccine than Muslims (adjusted odds ratio [aOR 1.4, 95% CI 0.97 - 2.01). Similarly, education was a statistically significant predictor of WTP, where respondents who had a graduate degree (aOR 1.98, 95% CI 1.14 - 3.45) and master's and MPhil or PhD degrees (aOR 1.93, 95% CI 1.07 - 3.480) had a higher WTP for the vaccine compared to the respondents who had no education. Division was a significant predictor of WTP, which showed that respondents from Chattogram (aOR 2.27, 95% CI 1.61 - 4.68), Dhaka (aOR 2.72, 95% CI 1.66 - 4.45), Khulna (aOR 3.37, 95% CI 1.84 - 6.17), Rajshahi (aOR 3.36, 95% CI 1.19 - 5.91), Rangpur (aOR 3.22, 95% CI 1.70 - 6.09), and Sylhet (aOR 4.70, 95% CI 2.53 - 8.74) had a higher WTP compared to respondents from Barisal. The findings also showed that WTP increased with an increasing level of knowledge about the COVID-19 vaccine (aOR 1.09, 95% CI 1.02 - 1.15), behavioral practices related to COVID-19 (aOR 1.11, 95% CI 1.06-1.17), higher subjective norm toward COVID-19 vaccine (aOR 1.25, 95% CI 1.08-1.46), and higher anticipated regret regarding being infected by COVID-19 (aOR 1.17, 95% CI 1.04-1.32). On the other hand, a more negative attitude toward the COVID-19 vaccine (aOR 0.91, 95% CI 0.88-0.95) was associated with decreased WTP.

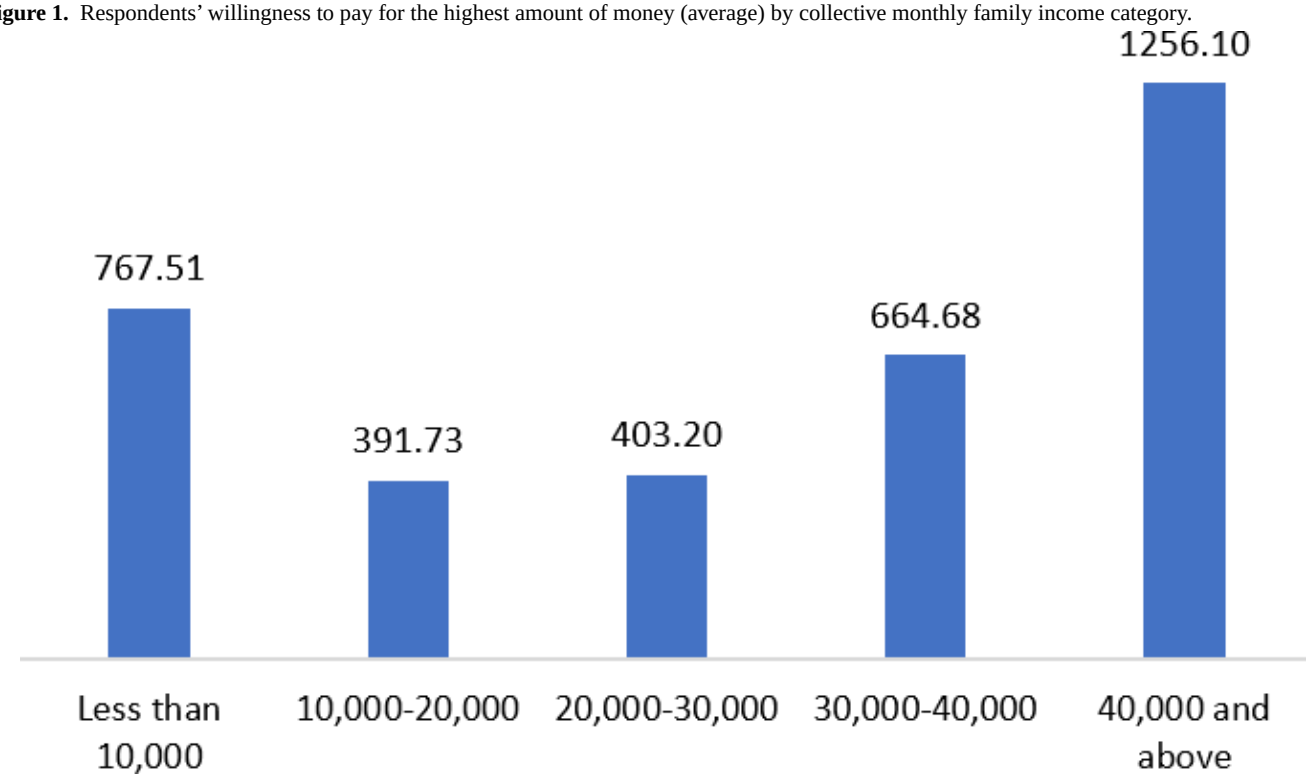
Although the logistic regression models initially included significant variables at the bivariate level, some variables lost their significance in the multivariate context. These included marital status, place of residence, occupation, income, and COVID-19 infection–related health status of the respondent, family, and friends. Compared to rural areas of residence, both urban and city corporations remain insignificant across the model. The respondent's family income, notably when it exceeded 10,000 BDT, was negligible in the adjusted models. Among the administrative divisions, the Mymensingh division showed insignificant effects across all models. The COVID-19 infection–related health status of the respondents, family, and friends showed an insignificant association with the outcome variable.

WTP the Highest Amount of Money

The mean and median WTP the highest amount of money among the respondents were 780.39 BDT and 300 BDT (IQR 150-500 BDT), respectively, for the COVID-19 vaccine. The WTP ranged from 1 BDT to 30,000 BDT.

Figure 1 illustrates the mean WTP for collective monthly family income. There was an increasing trend in WTP, accompanied by rising collective monthly family income, except for the lowest income category (less than 10,000 BDT). Respondents from the 10,000 - 20,000 BDT income group had a mean WTP of 391.73 BDT. On the other hand, the mean WTP of the group with an income of 40,000 BDT and above was 1256.10 BDT.

Figure 1. Respondents' willingness to pay for the highest amount of money (average) by collective monthly family income category.



Discussion

Principal Findings and Comparison to Prior Work

A successful vaccination program depends not only on safe and effective vaccines but also on vaccine hesitancy and WTP. The objective of this study was to assess the prevalence of WTP for the COVID-19 vaccine and its correlates. The study found that 51.6% (n=772) of participants were willing to pay for the vaccine, with an average WTP of 780.49 BDT and a median of 300 BDT (IQR 150-500 BDT). Our findings reveal a lower prevalence and WTP compared to a previous study conducted in Bangladesh. A survey conducted earlier reported that a higher proportion of respondents (68.4%) were willing to pay for the vaccine, along with a higher median WTP of US $7.08 [21]. This difference may be attributed to differences in the sampling methodologies. An earlier study collected web-based data, which likely led to an overrepresentation of individuals who are more educated, financially better off, and digitally literate—characteristics associated with higher WTP. By contrast, our study used both web-based and offline (face-to-face) data collection methods, thereby capturing a more socioeconomically and educationally diverse sample, potentially leading to a more representative estimate of WTP in the general population.

Similarly, there were also relatively modest mean and median values among respondents from low- and middle-income countries who were willing to pay for the COVID-19 vaccine (mean WTP of US $5 in Ethiopia and US $15 for an 80% effective vaccination in Iran) [26]. Another study conducted in Indonesia revealed that the WTP for a COVID-19 vaccine booster dose in this low- and middle-income country remains relatively low, with only 66.2% of the respondents expressing WTP. Among them, the majority (63.5%) indicated a WTP within the range of US $6.71 - $33.57) [3]. As concerns about vaccine resistance and budgetary limitations create the expectation that the government will fully subsidize the vaccine, this expectation contributes to a lower willingness to accept and pay for the vaccine. However, the lower prevalence of WTP is a concern for attaining herd immunity, as it is estimated that 60% of the population must be immunized under a vaccination system to ensure the effectiveness of any vaccine [3].

Our findings suggest that religion has a significant impact on the WTP for the COVID-19 vaccine in Bangladesh. Muslims had a significantly lower WTP than other religions (Hindu, Christian, and Buddhist), and the existing literature supports these findings [3]. The Muslim community is known for its fatalistic view of health and its appreciation, acceptance, and patience regarding their current situation [27,28]. The majority of the population in Bangladesh is Muslim. Muslims may hold various religion-related beliefs, have lower perceptions and levels of knowledge about health and vaccines, and exhibit trust issues toward vaccines developed by non-Muslim countries. These factors can influence their decision to purchase a vaccine and contribute to a lower WTP for the COVID-19 vaccine [29-32].

Our study demonstrates that education has a significant predictive value for WTP for the COVID-19 vaccine [19,20]. In our study, respondents with a graduate and master's or MPhil or PhD-level of education had a higher WTP than those with no education. Educated individuals are more likely to be conscious of their health and more willing to prevent diseases through preventive care such as vaccination [26,33,34]. Education also influences knowledge of health and vaccines, and individuals with a low level of expertise are less likely to develop a WTP for vaccines [34,35]. Educated individuals are

more likely to have higher incomes, which gives them greater purchasing power.

In our analysis, household income was found to have a significant positive association with WTP for the COVID-19 vaccine, particularly among higher-income groups. However, this association became statistically nonsignificant after adjusting for behavioral practices related to COVID-19 prevention. This suggests that behavioral practices may mediate the relationship between income and WTP. In our study sample, individuals with higher incomes were more likely to engage in protective behaviors against COVID-19. Consequently, the independent effect of income on WTP appeared to be mediated or absorbed by these behavioral practices, leading to the observed nonsignificant association after adjustment. However, income has proven to be one of the most influential factors in determining WTP for both COVID-19 and non–COVID-19 vaccines [8,18-20].

The study findings revealed a divisional disparity in WTP for the COVID-19 vaccine in Bangladesh. These differences may have occurred because of the socioeconomic and demographic differences among these divisions [36]. Barisal and Mymensingh are the 2 divisions with the highest poverty headcount ratios, whereas Sylhet, Dhaka, Khulna, and Chattogram have significantly lower poverty headcount ratios [37]. Our findings show that Sylhet has a 4-fold higher WTP, which may be attributed to economic and cultural factors. The household income and financial capacity of residents in Sylhet are high because of the strong remittance flow from the United Kingdom and the Middle East, which enhances their ability to pay for health services [38]. Additionally, the distinct cultural norms of the people of Sylhet may contribute to a higher WTP, as studies suggest that perceived norms and the personal environment are positively related to vaccination intention [38]. Households in Barisal and Mymensingh, with comparatively low incomes, may have a lower WTP for the COVID-19 vaccine. Our analysis reveals diverse WTP across divisions; however, a further qualitative study is essential to examine the behavioral and socioeconomic factors that underlie these disparities in WTP.

Apart from socioeconomic and demographic status, knowledge about the COVID-19 vaccine and preventive behavioral practices was associated with higher WTP, and the literature supports our study's findings [7,39]. Increased knowledge about the COVID-19 vaccine indicates a more reliable understanding of its safety and effectiveness, which in turn improves trust in the vaccine; thus, WTP may increase. Again, people with higher preventive behavioral practices tend to be more educated and health-conscious, so they are expected to have a higher WTP toward any vaccine [40].

The study findings indicate that components of TPB are statistically significant predictors of WTP. The main argument of TPB is that behavioral intention is the most important determinant of health behavior, specifically in the case of WTP [41]. Our findings revealed that respondents with a more negative attitude toward vaccines also had a lower WTP. Attitude acts as a personal evaluation of a behavior; as a result, the intention to pay decreases with a more negative attitude [42]. Similarly, perceived behavioral control suggests that performing a health behavior is not solely within the respondent's control; in this case, it is challenging to register. In our study, difficulty in web-based registration acts as a structural barrier that reduces WTP for the COVID-19 vaccine. On the other hand, our findings showed that subjective norms regarding COVID-19 vaccination and anticipated regret of contracting COVID-19 were positively associated with WTP for the vaccine. In our study, respondents whose family members approved of their decision to take the vaccine and highly regretted contracting COVID-19 had a higher WTP. This is also a part of subjective norms, as it provides permission from family members or friends for health behavior [41]. Thus, intention toward health behavior related to a disease or vaccine is driven by attitudes toward the disease or vaccine, subjective norms, perceived behavioral control, and other factors [41,43].

Strengths and Limitations

This study aimed to assess the prevalence and correlates of WTP for the COVID-19 vaccine in Bangladesh, which can help the GoB and policy makers promote a successful vaccination program on a larger scale for the general population by addressing economic challenges. However, some limitations of this study should be considered prior to the present findings. Our study used a cross-sectional design, which is limited in its ability to generate causal inferences because of temporal issues. Nonprobability sampling was used to reach the study population. We collected self-reported data on health status and other sociodemographic variables, which may have been subject to recall bias. Furthermore, we collected data using both web-based and face-to-face methods. We acknowledge the potential for selection bias that may have resulted from these data-collection methods. Web-based respondents were more likely to be younger, urban, educated, and technology-savvy because of the digital nature of the data-collection platform, which may have ultimately resulted in an overrepresentation of these groups. To mitigate this, the majority of the data (n=1022, 68.3%) were collected through face-to-face interviews to ensure the representativeness of the sample and to determine the population's national representation in terms of age, sex, residence, division, and marital status. However, it cannot be represented in terms of education, occupation, or income status. Finally, we acknowledge that WTP for a vaccine is context-dependent. Our study's results may be influenced by the unique sociodemographic and cultural dynamics that emerged during the data collection.

Conclusions

The study findings suggest that the government should introduce targeted educational campaigns aimed at specific demographics, such as religious communities, less educated groups, or those with lower incomes, to address their lower WTP for the COVID-19 vaccine. These campaigns need to be culturally appropriate to increase WTP. Health promotion materials and awareness campaigns as part of the behavior change communication program should be developed to increase knowledge about the COVID-19 vaccine. It will also increase preventive behavioral practices and reduce negative attitudes toward vaccines and vaccine-related conspiracies. Here, mass

media can be an effective platform to circulate accurate messages of the COVID-19 vaccine, and community leaders, along with religious leaders, can also be incorporated to mitigate religion-related mistrust and misconceptions. Policy makers should reconsider the web-based registration procedure for vaccine uptake, as it poses a structural barrier to WTP for the COVID-19 vaccine. Based on our findings, an easy alternative system should be introduced for the mass population to achieve the sustainability of the vaccination program. The government may offer small incentives to those who choose vaccination, which will be particularly helpful for lower-income groups. We suggest that policy makers consider a subsidization program that considers socioeconomic stratification, with a focus on highlighting lower-income groups to mitigate the catastrophic income challenge associated with WTP for the COVID-19 vaccine. Otherwise, a reasonable price should be fixed so that the COVID-19 vaccine is affordable. This will help to achieve the highest vaccine coverage and run a successful vaccination program without economic hardship.

Acknowledgments

The authors would like to thank the participants in this study and the data collectors for their contributions during the COVID-19 pandemic. This research did not receive any funding from any sources.

Data Availability

The authors will provide raw data supporting the conclusions of this article upon request.

Authors' Contributions

The study was conceptualized by MBH, MZA, MSI, SS, MMF, SR, MAH, AAM, and A-A-M, who collectively developed the research idea and framework. Data curation was undertaken by MBH, MZA, SS, MMF, SR, MAH, and AAM, ensuring accurate and organized data management. MBH, MZA, and AAM performed the formal analysis. The investigation was conducted by MBH, MZA, MSI, SS, MMF, SR, MAH, and AAM, who were actively involved in the research activities and data collection. Methodological development was collaboratively carried out by MBH, MZA, MSI, SS, MMF, SR, MAH, AAM, and A-A-M. Project administration duties were managed by MBH, MZA, MSI, SS, MMF, SR, MAH, and AAM. MBH ensured that resources were available to support the team with the necessary materials and infrastructure. Software-related tasks were handled by MZA and AAM, who contributed to coding and computational analyses. Supervision of the research process was provided by MBH and MZA, ensuring quality control and guidance throughout the study. Validation and visualization were undertaken by MBH, MZA, and AAM, who confirmed the accuracy of the results and presented them effectively. The original draft of the manuscript was written by AAM and SR. All authors—MBH, MZA, MSI, SS, MMF, SR, MAH, AAM, and A-A-M—participated in reviewing and editing the manuscript to improve its clarity and scholarly quality.

Conflicts of Interest

None declared.

References

1. Wang J, Lyu Y, Zhang H, et al. Willingness to pay and financing preferences for COVID-19 vaccination in China. Vaccine (Auckl) 2021 Apr 1;39(14):1968-1976. [doi: 10.1016/j.vaccine.2021.02.060] [Medline: 33714653]
2. Hajj Hussein I, Chams N, Chams S, et al. Vaccines through centuries: major cornerstones of global health. Front Public Health 2015;3:1-16. [doi: 10.3389/fpubh.2015.00269]
3. Harapan H, Wagner AL, Yufika A, et al. Willingness-to-pay for a COVID-19 vaccine and its associated determinants in Indonesia. Hum Vaccin Immunother 2020 Dec 1;16(12):3074-3080. [doi: 10.1080/21645515.2020.1819741] [Medline: 32991230]
4. Neumann-Böhme S, Varghese NE, Sabat I, et al. Once we have it, will we use it? A European survey on willingness to be vaccinated against COVID-19. Eur J Health Econ 2020 Sep;21(7):977-982. [doi: 10.1007/s10198-020-01208-6] [Medline: 32591957]
5. Huangfu L, Mo Y, Zhang P, Zeng DD, He S. COVID-19 vaccine tweets after vaccine rollout: sentiment-based topic modeling. J Med Internet Res 2022 Feb 8;24(2):e31726. [doi: 10.2196/31726] [Medline: 34783665]
6. Hou Z, Tong Y, Du F, et al. Assessing COVID-19 vaccine hesitancy, confidence, and public engagement: a global social listening study. J Med Internet Res 2021 Jun 11;23(6):e27632. [doi: 10.2196/27632] [Medline: 34061757]
7. García LY, Cerda AA. Contingent assessment of the COVID-19 vaccine. Vaccine (Auckl) 2020 Jul 22;38(34):5424-5429. [doi: 10.1016/j.vaccine.2020.06.068] [Medline: 32620375]
8. Sarker AR, Islam Z, Sultana M, et al. Willingness to pay for oral cholera vaccines in urban Bangladesh. PLoS ONE 2020;15(4):e0232600. [doi: 10.1371/journal.pone.0232600] [Medline: 32353086]
9. Zhang KC, Fang Y, Cao H, et al. Behavioral intention to receive a COVID-19 vaccination among Chinese factory workers: cross-sectional online survey. J Med Internet Res 2021 Mar 9;23(3):e24673. [doi: 10.2196/24673] [Medline: 33646966]

10. Kim SY, Sagiraju HKR, Russell LB, et al. Willingness-to-pay for vaccines in low- and middle-income countries: a systematic review. Ann Vaccines Immun 2014;1:1001. [doi: 10.47739/2378-9379/1001]
11. Lin Y, Hu Z, Zhao Q, Alias H, Danaee M, Wong LP. Understanding COVID-19 vaccine demand and hesitancy: a nationwide online survey in China. PLoS Negl Trop Dis 2020 Dec;14(12):e0008961. [doi: 10.1371/journal.pntd.0008961] [Medline: 33332359]
12. Wong LP, Alias H, Wong PF, Lee HY, AbuBakar S. The use of the health belief model to assess predictors of intent to receive the COVID-19 vaccine and willingness to pay. Hum Vaccin Immunother 2020 Sep 1;16(9):2204-2214. [doi: 10.1080/21645515.2020.1790279] [Medline: 32730103]
13. McQuestion M, Gnawali D, Kamara C, et al. Creating sustainable financing and support for immunization programs in fifteen developing countries. Health Aff (Millwood) 2011 Jun;30(6):1134-1140. [doi: 10.1377/hlthaff.2011.0265] [Medline: 21653967]
14. Wang H, Torres LV, Travis P. Financial protection analysis in eight countries in the WHO South-East Asia Region. Bull World Health Organ 2018 Sep 1;96(9):610-620E. [doi: 10.2471/BLT.18.209858] [Medline: 30262942]
15. Final report: household income and expenditure survey HIES 2022. Bangladesh Bureau of Statistics. 2023. URL: https://bbs.portal.gov.bd/sites/default/files/files/bbs.portal.gov.bd/page/b343a8b4_956b_45ca_872f_4cf9b2f1a6e0/2023-12-28-14-40-ac2b3d298f569f155a80871a49b7dd9e.pdf [accessed 2025-07-31]
16. Ahmed S, Hoque ME, Sarker AR, et al. Willingness-to-pay for community-based health insurance among informal workers in urban Bangladesh. PLoS ONE 2016;11(2):e0148211. [doi: 10.1371/journal.pone.0148211] [Medline: 26828935]
17. Shariful Islam SM, Lechner A, Ferrari U, Seissler J, Holle R, Niessen LW. Mobile phone use and willingness to pay for SMS for diabetes in Bangladesh. J Public Health (Oxf) 2016 Mar;38(1):163-169. [doi: 10.1093/pubmed/fdv009] [Medline: 25687131]
18. Sallam M, Anwar S, Yufika A, et al. Willingness-to-pay for COVID-19 vaccine in ten low-middle-income countries in Asia, Africa and South America: a cross-sectional study. Narra J 2022 Apr;2(1):e74. [doi: 10.52225/narra.v2i1.74] [Medline: 38450393]
19. Mudatsir M, Anwar S, Fajar JK, et al. Willingness-to-pay for a hypothetical Ebola vaccine in Indonesia: a cross-sectional study in Aceh. F1000Res 2020;8:1441. [doi: 10.12688/f1000research.20144.2]
20. Rajamoorthy Y, Radam A, Taib NM, et al. Willingness to pay for hepatitis B vaccination in Selangor, Malaysia: a cross-sectional household survey. PLoS ONE 2019;14(4):e0215125. [doi: 10.1371/journal.pone.0215125] [Medline: 30964934]
21. Kabir R, Mahmud I, Chowdhury MTH, et al. COVID-19 vaccination intent and willingness to pay in Bangladesh: a cross-sectional study. Vaccines (Basel) 2021 Apr 21;9(5):416. [doi: 10.3390/vaccines9050416] [Medline: 33919254]
22. Banik R, Islam MS, Pranta MUR, et al. Understanding the determinants of COVID-19 vaccination intention and willingness to pay: findings from a population-based survey in Bangladesh. BMC Infect Dis 2021 Aug 31;21(1):892. [doi: 10.1186/s12879-021-06406-y] [Medline: 34465297]
23. Earnshaw VA, Eaton LA, Kalichman SC, Brousseau NM, Hill EC, Fox AB. COVID-19 conspiracy beliefs, health behaviors, and policy support. Transl Behav Med 2020 Oct 8;10(4):850-856. [doi: 10.1093/tbm/ibaa090] [Medline: 32910819]
24. Hossain MB, Alam MZ, Islam MS, et al. Health belief model, theory of planned behavior, or psychological antecedents: what predicts COVID-19 vaccine hesitancy better among the Bangladeshi adults? Front Public Health 2021;9:711066. [doi: 10.3389/fpubh.2021.711066] [Medline: 34490193]
25. Ursachi G, Horodnic IA, Zait A. How reliable are measurement scales? External factors with indirect influence on reliability estimators. Proc Econ Finance 2015;20:679-686. [doi: 10.1016/S2212-5671(15)00123-9]
26. Zajacova A, Lawrence EM. The relationship between education and health: reducing disparities through a contextual approach. Annu Rev Public Health 2018 Apr 1;39:273-289. [doi: 10.1146/annurev-publhealth-031816-044628] [Medline: 29328865]
27. Elbarazi I, Devlin NJ, Katsaiti MS, Papadimitropoulos EA, Shah KK, Blair I. The effect of religion on the perception of health states among adults in the United Arab Emirates: a qualitative study. BMJ Open 2017 Oct 5;7(10):e016969. [doi: 10.1136/bmjopen-2017-016969] [Medline: 28982822]
28. Ali I. The COVID-19 pandemic: making sense of rumor and fear. Med Anthropol 2020 Jul;39(5):376-379. [doi: 10.1080/01459740.2020.1745481] [Medline: 32212931]
29. de Figueiredo A, Simas C, Karafillakis E, Paterson P, Larson HJ. Mapping global trends in vaccine confidence and investigating barriers to vaccine uptake: a large-scale retrospective temporal modelling study. Lancet 2020 Sep 26;396(10255):898-908. [doi: 10.1016/S0140-6736(20)31558-0] [Medline: 32919524]
30. Larson HJ, Schulz WS, Tucker JD, Smith DMD. Measuring vaccine confidence: introducing a global vaccine confidence index. PLOS Curr 2015 Feb 25;7:ecurrents.outbreaks.ce0f6177bc97332602a8e3fe7d7f7cc4. [doi: 10.1371/currents.outbreaks.ce0f6177bc97332602a8e3fe7d7f7cc4] [Medline: 25789200]
31. Sallam M, Dababseh D, Eid H, et al. High rates of COVID-19 vaccine hesitancy and its association with conspiracy beliefs: a study in Jordan and Kuwait among other Arab countries. Vaccines (Basel) 2021 Jan 12;9(1):42. [doi: 10.3390/vaccines9010042] [Medline: 33445581]

32. Sulaiman KDO. An assessment of Muslims reactions to the immunization of children in northern Nigeria. Med J Islamic World Acad Sci 2014;22(3):123-132. [doi: 10.12816/0008183]
33. Lu PJ, O'Halloran A, Kennedy ED, et al. Awareness among adults of vaccine-preventable diseases and recommended vaccinations, United States, 2015. Vaccine (Auckl) 2017 May 25;35(23):3104-3115. [doi: 10.1016/j.vaccine.2017.04.028] [Medline: 28457673]
34. Mora T, Trapero-Bertran M. The influence of education on the access to childhood immunization: the case of Spain. BMC Public Health 2018 Jul 18;18(1):893. [doi: 10.1186/s12889-018-5810-1] [Medline: 30021538]
35. Worasathit R, Wattana W, Okanurak K, Songthap A, Dhitavat J, Pitisuttithum P. Health education and factors influencing acceptance of and willingness to pay for influenza vaccination among older adults. BMC Geriatr 2015 Oct 26;15:136. [doi: 10.1186/s12877-015-0137-6] [Medline: 26503289]
36. Bangladesh sample vital statistics 2023. Bangladesh Bureau of Statistics. 2024. URL: https://bbs.portal.gov.bd/sites/default/files/files/bbs.portal.gov.bd/page/6a40a397_6ef7_48a3_80b3_78b8d1223e3f/2025-06-23-04-04-02e049844b479c3e811a22b3e3e7f744.pdf [accessed 2025-07-31]
37. Report on the household income and expenditure survey 2016. Bangladesh Bureau of Statistics. 2019. URL: http://203.112.218.101/storage/files/1/Publications/HIES/Final%20Report%20on%20HIES%202016.pdf [accessed 2025-07-31]
38. Geber S, Ho SS, Ou M. Communication, social norms, and the intention to get vaccinated against Covid-19: a cross-country study in Singapore and Switzerland. Eur J Heal Commun 2023;4:113-139. [doi: 10.47368/ejhc.2023.206]
39. Harapan H, Fajar JK, Sasmono RT, Kuch U. Dengue vaccine acceptance and willingness to pay. Hum Vaccin Immunother 2017 Apr 3;13(4):786-790. [doi: 10.1080/21645515.2016.1259045] [Medline: 27905832]
40. Hossain MB, Alam MZ, Islam MS, et al. Population-level preparedness about preventive practices against coronavirus disease 2019: a cross-sectional study among adults in Bangladesh. Front Public Health 2020;8:582701. [doi: 10.3389/fpubh.2020.582701] [Medline: 33505950]
41. Glanz K. Theory at a Glance: A Guide for Health Promotion Practice, 2nd edition: National Institutes of Health, National Cancer Institute; 2005.
42. Liu S, Liu J. Understanding behavioral intentions toward COVID-19 vaccines: theory-based content analysis of tweets. J Med Internet Res 2021 May 12;23(5):e28118. [doi: 10.2196/28118] [Medline: 33939625]
43. Ajzen I. The theory of planned behavior. Organ Behav Hum Decis Process 1991 Dec;50(2):179-211. [doi: 10.1016/0749-5978(91)90020-T]

Abbreviations

aOR: adjusted odds ratio
GoB: Government of Bangladesh
TPB: theory of planned behavior
WTP: willingness to pay

Edited by F Wu; submitted 09.12.24; peer-reviewed by E Hoque, E Kabir, J Vij; revised version received 05.06.25; accepted 19.06.25; published 15.08.25.

Please cite as:
Hossain MB, Alam MZ, Islam MS, Sultan S, Faysal MM, Rima S, Hossain MA, Mamun AA, Mamun AA
Willingness to Pay for the COVID-19 Vaccine and Its Correlates in Bangladesh: Cross-Sectional Study
JMIRx Med 2025;6:e69827
URL: https://xmed.jmir.org/2025/1/e69827
doi:10.2196/69827

Estimating Variance of Log Standardized Incidence Ratios Assessing Health Care Providers' Performance: Comparative Analysis Using Bayesian, Bootstrap, and Delta Method Approaches

Solomon Woldeyohannes[1,2], BSc, MPH, PhD; Yomei Jones[1], Dipl; Paul Lawton[1], MBBS, FRACP, PhD

[1]Menzies School of Health Research, Charles Darwin University, Northern Territory, Darwin, Casuarina, Australia

[2]School of Veterinary Sciences, University of Queensland, Gatton, Australia

Corresponding Author:
Solomon Woldeyohannes, BSc, MPH, PhD
Menzies School of Health Research, Charles Darwin University, Northern Territory, Darwin, Casuarina, Australia

Related Articles:

Companion article: https://www.medrxiv.org/content/10.1101/2025.04.22.25326183v1

Companion article: https://med.jmirx.org/2025/1/e83798

Companion article: https://med.jmirx.org/2025/1/e83796

Abstract

Background: In health care providers' performance assessment, standardized incidence ratios are essential tools used to assess whether observed event rates deviate from expected values. Accurate estimation of variance in these ratios is crucial as it affects decision-making regarding providers' performance. There is little data on how the choice of these variance estimation methods affects decision-making.

Objective: In this study, we compared 3 methods (the delta method, bootstrapping method, and Bayesian approach) to estimate the variance of the logarithm of the standardized incidence ratio.

Methods: Using patient-level data from the Australia and New Zealand Dialysis and Transplant Registry for 2012 - 2023, we used a random effects model to predict treatment at home 1 year after starting treatment. We compared the 3 approaches (with more than 5000 iterations for bootstrapping and Markov chain Monte Carlo sampling) using bias, variance, and mean squared error (MSE) as performance measures. Using the 3 methods, funnel plots were used to compare the hospitals' performance in treating Indigenous and non-Indigenous patients close to home, as a service-level measure of equity.

Results: The bias values across all methods were similar, with the Bayesian method narrowly having the lowest bias (0.01922), followed by the delta method (0.01927) and bootstrap method (0.02567). In addition, the Bayesian method exhibited the lowest variance (0.00005), indicating more stable and less dispersed estimates. The delta method had a higher variance (0.00016), while the bootstrap method had the highest variance (0.00027), meaning it introduced more uncertainty. Finally, the Bayesian method had the lowest MSE (0.00042), indicating better overall accuracy, while the bootstrap method had the highest MSE (0.00094), showing it was the least reliable method.

Conclusions: We demonstrated that these methods can be used to measure equity for patient-centered outcomes, both within and between service providers simultaneously. The choice of variance estimation method is critical and heavily affects the interpretation of the performance of health service providers. We favor the Bayesian Markov chain Monte Carlo method as it was found to be a better approach.

Trial Registration: ANZCTR ACTRN12623001241628; https://www.anzctr.org.au/Trial/Registration/TrialReview.aspx?id=379101&isReview=true

(JMIRx Med 2025;6:e77415) doi:10.2196/77415

KEYWORDS

standardized incidence ratio; SIR; performance; health care provider; machine learning; equity

Introduction

Public scrutiny of health care service performance has been emphasized in the last two decades. For instance, the Australian government has introduced the National Health Reform in 2011 [1] and recently the 2020 - 2025 National Health Reform Agreement [2]. This, in turn, has led to increased attention to institutional comparisons based on quantitative outcome measures such as standardized mortality ratios (SMRs) in which, with the aid of CIs, "outlying" institutions are identified [3]. A league table of hospitals based on mortality [4] and Shewhart's control charts (using 1, 2, and 3 SD limits) [5] has been proposed and criticized to compare institutional ranking. More recently, a "funnel plot," in which an estimate of an underlying quantity is plotted against an interpretable measure of its precision, has become a useful graphical aid for institutional comparisons [3,6,7]. Although funnel plots have been used in meta-analyses, in particular to detect publication bias, they have recently been strongly recommended as the most appropriate way to display performance indicators such as comparisons of risk-adjusted rates between health care units [8]. SMRs are the commonly used performance index for institutional comparisons [9]. However, this concept has been readily extended to encompass several other indices such as age-standardized relative survival and excess hazard ratios [8] and standardized incidence ratio (SIR) [10]. Estimating the variance of log SIR (denoted by Log-SIR hereafter) is necessary for creating false discovery rates (FDRs) in studies that use funnel plots for assessing centers'/hospitals' performance. Accurate estimation of variance in these ratios is crucial as it affects decision-making regarding hospital performance and quality improvement strategies. Despite different variance estimation methods being used widely in application, there are no data on how the choice of these methods affects the assessment of performance. In this study, we compared 3 methods, namely, delta method, bootstrapping, and Bayesian approaches, to estimate the variance of the Log-SIR and subsequent funnel plot approaches to build FDRs for the Log-SIR.

The delta method is the analytical approach to estimate the variance of the logarithm of SMR, denoted by Log-SMR hereafter. It approximates the variance of a function of random variables by using the Jacobian matrix and the covariance matrix of the original variables [11].

Quaresma et al [12] used the delta method directly to estimate risk-adjusted excess hazard ratios as a performance measure in a study of population-based cancer survival. Also, Powell [13] applied the delta method to approximate the variance of demographic parameters in avian biology studies. Vasilevskis et al [9] used CIs for comparing SMR using a bootstrapping approach in a study involving prediction of 30-day intensive care unit mortality. Though CIs can be constructed for SMR directly, Hosmer and Lemeshow [14] demonstrated CIs with good coverage for the logarithm of the SMR. Also, Austin [15] investigated 4 bootstrap procedures for estimating CIs for predicted-to-expected ratios in a hospital profiling study. They indicated that existing bootstrap procedures should not be used to compute CIs for predicted-to-expected ratios when conducting provider profiling.

Like bootstrapping, a Bayesian approach via Markov chain Monte Carlo can be used to approximate this variance. For instance, Ventrucci et al [16] applied Bayesian hierarchical models to estimate small area level SMR and constructed FDRs in a study of liver cancer morbidity cases recorded between 1998 and 2003 in Emilia-Romagna municipalities. In addition, Sukul et al [17] demonstrated the application of a Bayesian hierarchical model in assessing hospital and operator variation in cardiac rehabilitation referral and participation after percutaneous coronary intervention using a retrospective observational cohort of patients who underwent percutaneous coronary intervention at 48 nonfederal Michigan hospitals between January 1, 2012, and March 31, 2018.

Applying the delta method depends on fulfillment of underlying distributional assumption, asymptotic normality. Bootstrapping, on the contrary, has the advantage of not relying on distributional assumptions and can be used to directly estimate the distribution of Log-SIR or Log-SMR. This can lead to more robust variance estimates, particularly in settings with small sample sizes or unknown distributions. By resampling, bootstrapping accounts for sampling variability and can help improve the precision of performance assessments [11]. Therefore, this study compares the 3 variance estimation methods using bias, variance, and mean squared error (MSE) as measures of performance.

Methods

Motivating Idea

For more than 25 years, First Nations health organizations and patients in rural and remote Australia have persistently called for more responsive treatment, closer to home, for First Nations people with end-stage kidney disease [18,19]. Community-led advocacy groups have continued this call in more recent years. A national meeting of First Nations patients with kidney failure in September 2017 renewed this message [20]. Over the last 15 years, substantial progress has been made in expanding and decentralizing hemodialysis care across remote Australia [21]. Nevertheless, most treatment is still provided as hemodialysis in nurse-facilitated centers in major or regional towns, rather than at home in remote communities [22].

The Return to Country Study, of which this methodological work is a part, aims to characterize the socioeconomic, environmental, health service, and biomedical factors driving the health outcomes and patterns of health service utilization experienced by First Nations Australians receiving kidney replacement therapy and investigate whether health service changes to address these identified barriers can achieve higher rates of kidney replacement therapy closer to home [23].

Data Source and Management

The source of data for our motivating example is the Australia and New Zealand Dialysis and Transplant Registry (ANZDATA) [6,22]. ANZDATA receives, collates, and analyzes data from centers providing care for patients receiving long-term dialysis or kidney transplantation in Australia and New Zealand. Data submission is voluntary but complete. For this methodological study, we used the data extract provided

by ANZDATA for the Return To Country Study (ANZRREQ-471) [23].

We received n=55,856 patient data on the course of treatments and patients' history data from February 14, 1992, till December 31, 2023. Since our initial study period was defined from January 1, 2005, to December 31, 2023, we excluded patient data before January 01, 2005. This resulted in n=46,160 observations on the course of treatment and comorbidities data. With the revised study period definition (January 1, 2012–December 31, 2023), following consultation with a team of chief investigators, a total of 11,586 observations were excluded (n=44,270 individual level observations were retained out of 55,856). Due to 1743 missing observations for late referral, 808 on weight, and 188 on height variables, n=41,531 patient data were retained. In addition, for comparison purposes, centers were split into Indigenous and non-Indigenous centers. Some centers had fewer than 20 Indigenous patients. This required considering an adequate count of Indigenous patients per center for running the hierarchical logistic regression. Accordingly, centers with fewer than 20 Indigenous patients were excluded, which resulted in n=16,243 (25,288 observations deleted) individual-level data. Moreover, we dropped patients with missing postcode (2640 observations deleted), a total of n=13,603 remained. Finally, among the 13,603 observations, 3309 observations had censored status and hence were excluded. In addition, we excluded 55 missing observations on lung diseases, cardiovascular disease, and diabetes combined. Therefore, a total of 10,195 observations were included in our study.

In the following, we presented model specification, the derivation of the variance for the LogSIR using the delta method and a description of the bootstrap and Bayesian approaches for estimating variance of Log-SIR.

Model Specification and Likelihood Definition

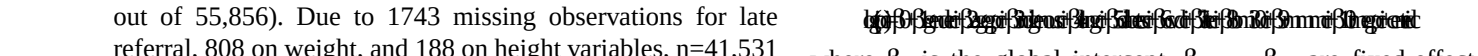
Since we have a binary outcome of receiving treatment close to home for end-stage kidney disease, denoted by y_{ci}, from n_c number of patients receiving treatment from center c for N centers, we proposed a Bernoulli sampling distribution for the probability of getting treatment close to home for the i^{th} patient from center c. That is, $y_{ci} \sim$ Bernoulli(p_{ci}) and a random effects logistic regression model can be specified as:

$$\text{logit}(pci)=\eta ci=\beta 0+\beta 1X1ci+\ldots+\beta kXkci+uc \quad (1)$$

where y_{ci} is the binary outcome for patient i in center c, $X_{1ci,}$ ... X_{kci} are k covariates for patient i in center c, β_0, β_1, ..., β_k are fixed effects, u_c is the random effect for center c, assumed to be normally distributed: uc~N0,σc2, and $p_{ci} = P(y_{ci}=1)$.

We included the following covariates in our model: gender, age group, Indigenous status, lung disease, diabetes, BMI, cardiovascular disease, referral status, remoteness, and time period. And they were coded as follows: gender (male vs female categories), agegp (age group with 7 categories: ≥16 - 26, ≥26 - 36, ≥36 - 46, ≥46 - 56, ≥56 - 66, ≥66 - 76, and ≥76), Indigenous status (Indigenous vs non-Indigenous), lung (lung disease status: yes vs no), diabetes (diabetes status: yes vs no), late (late referral status: yes vs no), bmi30 (binary BMI status: BMI <30 kg/m^2 vs BMI ≥30 kg/m^2), mmm (Modified Monash Model remoteness scale with 7 categories: metropolitan areas [MM1], regional centers [MM2], large rural towns [MM3], medium rural towns [MM4], small rural towns [MM5], remote communities [MM6], and very remote communities [MM7]), and timegp (time periods: 2012 - 2015, 2016 - 2019, and 2020 - 2023).

Accordingly, given y_{ci} binary "Return to Country" outcome for individual i in center c, which is distributed as $y_{ci} \sim$ Bernoulli(p_{ci}), then the logit of the probability p_{ci} is modeled as follows:



where β_0 is the global intercept, β_1, ..., β_{10} are fixed-effect coefficients for the covariates, $centreid_c \sim N(0,\sigma_u^2)$ is the group-level random intercept for center c, and p_{ci}=Pr (y_{ci}=1 | covariates).

Since we have individual-level data, we fitted a binary logistic regression model and computed the Log-SIR by aggregating: (1) the observed binary "Return to Home" status in center c and (2) the model-based predicted probabilities (used to calculate the expected number of patients returning home in center c).

Then, the Log-SIR is computed as:

$$\text{Log-SIRc}=\sum i\in \text{cyi}\sum i\in \text{cp}\hat{}i$$

where $y_i \in \{0,1\}$ is the observed outcome for individual i, and $\hat{p}_i$ is the predicted probability of receiving treatment close to home for individual patient i from center c.

This approach is methodologically valid and commonly used in Bayesian hierarchical modeling and disease mapping, especially when individual-level data are available, but aggregate counts are not directly observed. Modeling binary outcomes using Bernoulli likelihoods (ie, logistic regression) is appropriate for estimating probabilities of outcome conditional on covariates. These estimated probabilities can then be summed within groups to yield expected counts for computing SIR or relative risks. This technique allows the derivation of SIR from model-based expected counts, which is consistent with the definition of indirect standardization [14,24-27]. Further details of the model specification can be found in Multimedia Appendix 1.

Application works using this approach include Kasza et al [28] and Normand et al [29]. Application of random intercept multilevel logistic regression models to indirectly standardize performance measures is explored by Clark and Moore [30] using National Trauma Data Bank data for the admission year 2008. Yang et al [31] explored hierarchical logistic regression (LR) modeling under various conditions applying Bayesian and frequentist methods.

Delta Method for the Variance of the Log-SIR

The delta method is a technique used to approximate the variance of a function of 1 or more random variables [32-34]. The first-order Taylor series approximation for moments of ratio estimators is used to derive the mean and variance estimates; see Casella and Berger [32] (pages 244 - 245). In

the context of estimating the variance of the Log-SIR, we can apply the delta method to approximate the variance of log OcEc. It approximates the variance of a function of random variables by using the Jacobian matrix and the covariance matrix of the original variables; see Boos and Stefanski [35] (page 14). Accordingly, the variance of Log-SIR$_c$ is approximated by:

VarLog-SIRc≈∇g·CovOc,Ec·∇gT (2)

where the covariance matrix of O_c and E_c is specified as:

Cov (Oc,Ec)=(Var (Oc)Cov (Oc,Ec)Cov (Oc,Ec)Var (Ec))

And the Jacobian matrix (gradient) ∇g of the function $g(O_c,E_c)$ with respect to O_c and E_c is given by: ∇g=1Oc-1Ec

Substituting ∇g and Cov(O_c,E_c) into the formula, we get the final expression for the variance:

Var(Log−SIRc)≈Var(Oc)Oc2+Var(Ec)Ec2−2·Cov(Oc,Ec)OcEc (3)

Detailed derivation of the final formula for the variance of log(SIR) using the delta method given the model specification and the likelihood formulations above is presented in Multimedia Appendix 2.

The next section summarizes the estimates for Var(O_c), Var(E$_c$), and Cov(O$_c$,E$_c$).

Variance of O$_c$: Var(O$_c$)

Let Y_i be the binary outcome for individual i in center c. The observed incidence O_c is the sum of binary outcomes Y_i for individuals within the c^{th} center. If patients share hospital-level characteristics, the outcomes Y_i are not independent but are correlated due to the shared random effect. The observed counts for center c are:

Oc=∑i∈ ncYi

The variance of O_c is given by:

VarOc=Var∑i∈ ncYi

Using the property of variance for the sum of random variables, this expands to:

VarOc=∑i∈ ncVarYi+2∑i§amp;lt;j∈ ncCovYi,Yj

This expression is derived from the formula for the variance of the sum of random variables. Here, Var(Y_i) represents the variance of the individual observations, and Cov(Y_i,Y_j) is the covariance between pairs of observations. The factor of 2 in front of the covariance term accounts for the fact that each covariance term is counted only once when summing over pairs $i<j$.

For a logistic regression model with random intercepts, the variance and covariance terms are as follows:

Var(Y_i)=p_i(1 − p_i) (4)

CovYi,Yj=pi1-pipj1-pjσu2 (5)

Thus, Var(Oc)=∑i∈ ncpi(1−pi)+2∑i<j∈ ncpi(1−pi)pj(1−pj)σu2 (6)

Derivation of Var(E)

The expected counts E are the sum of predicted probabilities p_i for individuals within a center. The variance of E arises from the uncertainty in the predicted probabilities due to the random effects.

The expected counts for center c are:

Ec=∑i∈ ncpi

The variance of E_c is:

VarEc=∑i∈ ncVarpi+2∑i§amp;lt;j∈ ncCovpi,pj

For the random-effects logistic regression model:

Varpi≈pi1-pi2Varηi

where ηi=xiTβ+uc is the linear predictor. The covariance between p_i and p_j (for $i\neq j$) is as follows:

Cov(pi,pj)≈[pi(1−pi)][pj(1−pj)]Cov(ηi,ηj)

Since η_i and η_j share the same random effect u_c:

Covηi,ηj=σu2

Thus:

Covpi,pj≈pi1-pipj1-pjσu2

Combining these results:

[Var(Ec)=∑i∈ nc[pi(1−pi)]2Var(ηi)+2∑i<j∈ nc[pi(1−pi)][pj(1−pj)]σu2]

Derivation of Cov(O$_c$, E$_c$)

The covariance between O_c and E_c, where O_c is the observed count and E_c is the expected count for center c, arises because both depend on the same underlying probabilities p_i, which are influenced by the shared random effect.

To derive the covariance Cov(O_c, E_c), given (Oc=∑i∈ ncYi) (Observed count) and (Ec=∑i∈ ncpi) (Expected count), we have the covariance between O_c and E_c defined as:

CovOc,Ec=Cov∑i∈ ncYi,∑i∈ ncpi

And using the property of covariance for sums, we get:

Cov(Oc,Ec)=∑i∈ ncCov(Yi,pi)+2∑i<j∈ ncCov(Yi,pj)

Therefore, the final expression of Cov(O_c,E_c) becomes :

CovOc,Ec=∑i∈ ncpi1-piVarηi+2∑i§amp;lt;j∈ ncpi1-pipj1-pjσu2 (8)

Bootstrapping Approach

Commonly, the bootstrap approach is used to approximate variance of the log standardized incidence ratio. By sampling with replacement from the observed sample, creating a resampled dataset of size n and repeating this B times, it creates a nonparametric bootstrapped distribution [32], pages 479 - 480. This distribution can be used to estimate the variance of the Log-SIR$_c$. Mathematically, this can be summarized as:

σ^Boot2=1B−1∑b=1B(θ¯*−θ^b*)2

with θb*^ the Log-SIR$_c$ value estimated in the b^{th} bootstrap sample and θ*- the mean Log-SIR$_c$ estimated over the B bootstrap samples; here B=5000.

Bayesian Approach

Given the model specification given in (1), the posterior distribution for a random effects logistic regression model can be expressed in a hierarchical form, integrating over the random effects u_c. It can be recalled that the form of a posterior for hierarchical models is [35]:

$$\pi(\theta \mid Y=y)=f(y \mid \theta)\int\pi(\theta \mid \alpha)h(\alpha)d\alpha\iint f(y \mid \theta)\pi(\theta \mid \alpha)h(\alpha)d\alpha d\theta.$$

Using the likelihood for random effects logistic regression and priors for β and u_c, the full posterior distribution can be shown to be:

$$\theta\beta \mid y\propto\prod_{i=1}^{n}\int\left(\theta_{ij}\right)^{y_{ij}}\left(1-\theta_{ij}\right)^{1-y_{ij}}\times\exp\left(\beta_0+\beta_1x_{1ij}+\beta_2x_{2ij}+\beta_3x_{3ij}+u_i\right)\prod_{i=1}^{n}\frac{1}{\sqrt{2\pi\sigma_u^2}}\exp\left(-\frac{u_i^2}{2\sigma_u^2}\right)$$

Details of the derivation of the full posterior distribution are summarized in Multimedia Appendix 3.

Due to the need to integrate out the nuisance parameters in (9) and lack of conjugate priors, and the hierarchy involved, computing difficult integrals is required using MCMC methods whereby a dependent sequence of random variables is obtained with the property that in the limit these random variables have the posterior distribution.

Accordingly, the following information is used to estimate the variance of the Log-SIR using the Bayesian approach:

y_{ci} ~ Bernoulli(p_{ci})

$$\text{logit}(pci=P(yci=1))=\eta ci=\beta 0+\sum m=1k\beta mXmci+uc$$

where:

(β0,β1,…,βk~N(0,σc2)),(uc~N(0,σc2)),(σc2=1τ), and τ ~ Gamma(0.001,0.001).

The MCMC simulation is conducted using 25,500 iterations with 500 initial burn-ins, 3 chains, and a single thinning interval. Analysis was performed using the R Statistical Programming Language and the associated R packages [36-42].

Performance Metrics: Bias, Variance, and MSE

To compare the performance of the delta method, bootstrap, and MCMC approaches for estimating the variance of the Log-SIR, we evaluated several criteria such as bias (the difference between the expected value of the estimator and the true value), consistency (the estimator should converge to the true value as the sample size increases), and MSE (for overall accuracy).

Ethical Considerations

Ethical approval was obtained from the Human Research Ethics Committee (HREC) of the Northern Territory Department of Health and Menzies School of Health Research (2019 - 3530), Far North Queensland HREC (2023/QCH/99606 (Nov ver 4) - 1732), the Central Adelaide Local Health Network HREC (2023/HRE00209), the Aboriginal Health Council of South Australia (AHREC Protocol number 04-23-1078), the Aboriginal Health and Medical Research Council of New South Wales (AH&MRC HREC reference: 2230/24), and the Far North Queensland Human Research Ethics Committee (FNQ HREC reference: HREC/2023/QCH/99606 (Nov ver 4) - 1732). For information on informed consent details, please refer to our protocol paper on the “Return to Country” project, which can be accessed here [23].

Results

Variance of Log-SIR Using the 3 Estimation Methods

A summary of bias, along with variance and MSE, is shown in Table 1.

Table . Comparison of bias, variance, and mean squared error for different estimation methods.

Method	Bias	Variance	Mean squared error
Delta	0.01927454	1.696437e-04	0.0005411516
Bootstrap	0.02566281	2.771867e-04	0.0009357665
Bayesian	0.01922758	5.142122e-05	0.0004211210

The analysis result indicated that the bias values across all methods were similar, with MCMC slightly showing the lowest bias (0.01922), followed by the delta method (0.01927) and the bootstrap method (0.02567), respectively. This suggests that the Bayesian MCMC method provides a slightly less biased variance estimate of Log-SIR than the other methods. In addition, the Bayesian MCMC method exhibits the lowest variance (0.00005), indicating more stable and less dispersed estimates of the Log-SIR. Higher variance was observed in the delta method (0.00016), while the bootstrapping approach resulted in the highest variance (0.00027), introducing more uncertainty in the Log-SIR estimates. Looking at the overall accuracy of the methods, the Bayesian MCMC method had the lowest MSE (0.00042), indicating better overall accuracy. The delta method follows with an MSE of 0.00054, and the bootstrap method had the highest MSE (0.00094), showing it to be the least reliable method among the methods compared.

The result, in general, indicated lower values on bias, variance, and MSE values. Lower bias values indicated that the estimators are more accurate on average, lower variance indicated that the estimators are more consistent, and lower MSE indicated that the estimators are both accurate and consistent. However, the parameter estimates were the lowest for the MCMC method, indicating the Bayesian approach to be a more preferred approach for the estimation of the variance of the Log-SIR (var[Log-SIR]). MCMC is the best-performing method as it has the lowest bias, variance, and MSE. The delta method performs reasonably well but has slightly higher variance and MSE than MCMC. Bootstrap captures variability well but introduces more uncertainty, as seen in its high variance and MSE.

In addition, a comparison of the 3 methods in terms of consistency is shown in Figure 1. Accordingly, Figure 1 highlights the trade-offs among the variance estimation methods. While bootstrapping tends to be more variable, MCMC provides

more stable estimates, and the delta method offers computational efficiency but can be less precise. Bootstrapping (green) shows higher variance. The green points, representing bootstrap-based variance estimates, are often higher compared to the other 2 methods. This suggests that bootstrapping introduces additional variability, which is expected since it resamples the data and can exaggerate variance in small samples.

However, the Bayesian MCMC estimates (the blue points) are more stable. They are generally lower than bootstrapping but slightly higher than the delta method for most of the cases. The Bayesian methods incorporate prior information, and this leads to more stabilized variance estimates.

The delta method (red) is the most conservative and hence it often yields the lowest variance estimates. This method uses first-order approximations and may underestimate variance, especially for complex or skewed data distributions.

A summary table for each center is shown in Table 2. As is evident from Table 2, the standard errors were highly variable across centers using the bootstrap method followed by the delta method.

Figure 1. Delta method, bootstrapping, and Bayesian approaches.

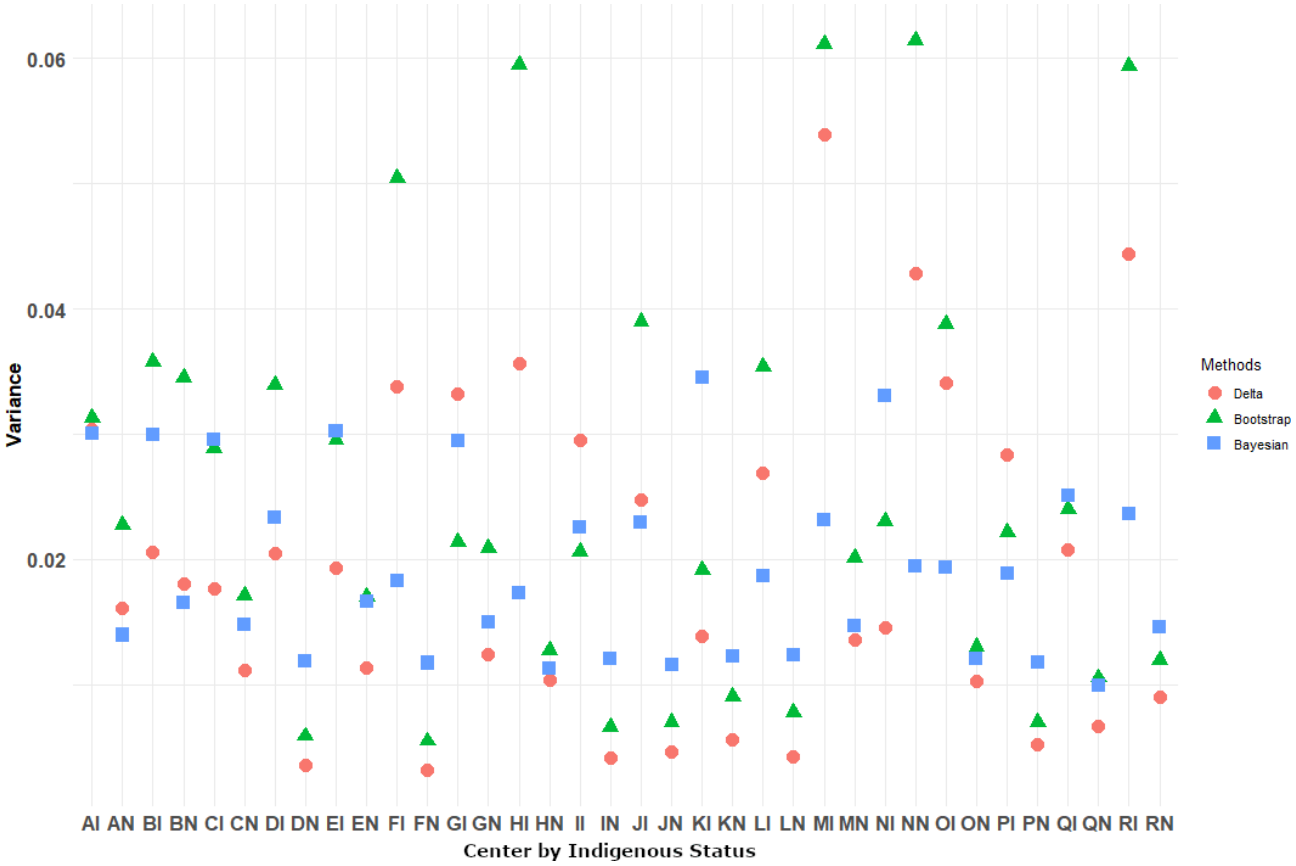


Table . Comparison of delta, bootstrap method, and Bayesian estimates along with 95% coverage by center.[a]

Center	Mean	SE	LCI[b]	UCI[c]	Mean	SE	LCI	UCI	Mean	SE	LCrI[d]	UCrI[e]
AI	0.100	0.030	0.041	0.159	0.016	0.031	−0.050	0.071	0.031	0.030	−0.019	0.098
AN	0.010	0.016	−0.021	0.041	0.028	0.023	−0.019	0.069	−0.017	0.014	−0.038	0.016
BI	−0.382	0.021	−0.423	−0.341	−0.453	0.036	−0.525	−0.386	−0.441	0.030	−0.491	−0.374
BN	−0.100	0.018	−0.135	−0.065	−0.092	0.035	−0.164	−0.027	−0.131	0.017	−0.156	−0.093
CI	−0.148	0.018	−0.183	−0.113	−0.220	0.029	−0.277	−0.165	−0.209	0.030	−0.258	−0.142
CN	−0.008	0.011	−0.030	0.014	0.009	0.017	−0.026	0.041	−0.034	0.015	−0.057	0.000
DI	−0.090	0.020	−0.129	−0.051	−0.130	0.034	−0.200	−0.068	−0.139	0.023	−0.177	−0.086
DN	−0.017	0.004	−0.025	−0.009	0.026	0.006	0.014	0.037	−0.025	0.012	−0.043	0.003
EI	−0.057	0.019	−0.094	−0.020	−0.135	0.030	−0.197	−0.080	−0.122	0.030	−0.172	−0.054
EN	−0.008	0.011	−0.030	0.014	0.000	0.017	−0.034	0.032	−0.038	0.017	−0.063	0.001
FI	−0.004	0.034	-0.071	0.063	-0.012	0.050	-0.125	0.073	-0.040	0.018	-0.069	0.002
FN	-0.021	0.003	-0.027	-0.015	0.025	0.005	0.015	0.036	-0.027	0.012	-0.044	0.001
GI	0.143	0.033	0.078	0.208	0.066	0.021	0.018	0.102	0.075	0.029	0.027	0.142
GN	−0.029	0.012	−0.053	−0.005	−0.014	0.021	−0.057	0.024	−0.057	0.015	−0.080	−0.022
HI	−0.038	0.036	−0.109	0.033	−0.044	0.060	−0.174	0.059	−0.075	0.017	−0.103	−0.035
HN	0.020	0.010	0.000	0.040	0.054	0.013	0.027	0.077	0.000	0.011	−0.017	0.027
II	0.103	0.029	0.046	0.160	0.067	0.021	0.021	0.101	0.053	0.023	0.017	0.104
IN	−0.008	0.004	−0.016	0.000	0.032	0.007	0.018	0.044	−0.019	0.012	−0.038	0.009
JI	−0.033	0.025	−0.082	0.016	−0.075	0.039	−0.157	−0.005	−0.084	0.023	−0.122	−0.033
JN	−0.002	0.005	−0.012	0.008	0.038	0.007	0.024	0.051	−0.015	0.012	−0.032	0.013
KI	0.015	0.014	−0.012	0.042	−0.085	0.019	−0.125	−0.048	−0.055	0.035	−0.113	0.021
KN	−0.015	0.006	−0.027	−0.003	0.020	0.009	0.002	0.038	−0.030	0.012	−0.049	−0.002
LI	0.019	0.027	−0.034	0.072	0.008	0.035	−0.068	0.069	−0.019	0.019	−0.049	0.023
LN	−0.032	0.004	−0.040	−0.024	0.008	0.008	−0.008	0.022	−0.043	0.012	−0.061	−0.013
MI	0.068	0.054	−0.038	0.174	0.028	0.061	−0.113	0.116	0.017	0.023	−0.021	0.070
MN	0.000	0.014	−0.027	0.027	0.015	0.020	−0.027	0.052	−0.028	0.015	−0.050	0.006
NI	−0.111	0.014	−0.138	−0.084	−0.195	0.023	−0.241	−0.153	−0.176	0.033	−0.230	−0.102
NN	0.011	0.043	−0.073	0.095	0.003	0.061	−0.139	0.099	−0.027	0.019	−0.057	0.018
OI	0.046	0.034	−0.021	0.113	0.032	0.039	−0.052	0.097	0.007	0.019	−0.024	0.051
ON	0.017	0.010	−0.003	0.037	0.048	0.013	0.020	0.072	−0.004	0.012	−0.022	0.025
PI	0.077	0.028	0.022	0.132	0.065	0.022	0.016	0.102	0.039	0.019	0.008	0.082
PN	0.010	0.005	0.000	0.020	0.047	0.007	0.033	0.060	−0.004	0.012	−0.022	0.023
QI	0.062	0.021	0.021	0.103	0.012	0.024	−0.040	0.055	0.008	0.025	−0.033	0.065
QN	−0.006	0.007	−0.020	0.008	0.037	0.011	0.015	0.056	−0.020	0.010	−0.035	0.003
RI	0.021	0.044	−0.065	0.107	−0.014	0.059	−0.151	0.085	−0.026	0.024	−0.064	0.027
RN	0.014	0.009	−0.004	0.032	0.035	0.012	0.010	0.057	−0.009	0.015	−0.031	0.025

[a]All units are on the natural log scale.

[b]LCI: 95% lower confidence limit.

[c]UCL: 95% upper confidence limit.

[d]LCrI: 95% lower credible interval.

[e]UCrI: 95% upper credible interval.

In summary, there are notable variations in variance estimates across centers. Some centers exhibit more spread between methods, suggesting that the choice of method affects variance estimates significantly.

Similarly, a summary table of false discovery rates (FDRs) for each center is shown in Table 3. It is evident that there are notable variations in FDR estimates across centers. Some centers exhibit more spread between methods, suggesting that the choice of method affects variance and hence the resulting coverage significantly.

Table . Comparison of delta, bootstrap, and Bayesian estimates along with 95% false discovery rates by center.

Center	Mean	SE	LFDR[a]	UFDR[b]	Mean	SE	LFDR	UFDR	Mean	SE	LFDR	UFDR
AI	0.100	0.030	−0.059	0.059	0.016	0.031	−0.061	0.061	0.031	0.030	−0.059	0.059
AN	0.010	0.016	−0.032	0.032	0.028	0.023	−0.045	0.045	−0.017	0.014	−0.027	0.027
BI	−0.382	0.021	−0.040	0.040	−0.453	0.036	−0.070	0.070	−0.441	0.030	−0.059	0.059
BN	−0.100	0.018	−0.035	0.035	−0.092	0.035	−0.068	0.068	−0.131	0.017	−0.032	0.032
CI	−0.148	0.018	−0.035	0.035	−0.220	0.029	−0.057	0.057	−0.209	0.030	−0.058	0.058
CN	−0.008	0.011	−0.022	0.022	0.009	0.017	−0.034	0.034	−0.034	0.015	−0.029	0.029
DI	−0.090	0.020	−0.040	0.040	−0.130	0.034	−0.067	0.067	−0.139	0.023	−0.046	0.046
DN	−0.017	0.004	−0.007	0.007	0.026	0.006	−0.012	0.012	−0.025	0.012	−0.023	0.023
EI	−0.057	0.019	−0.038	0.038	−0.135	0.030	−0.058	0.058	−0.122	0.030	−0.059	0.059
EN	−0.008	0.011	−0.022	0.022	0.000	0.017	−0.033	0.033	−0.038	0.017	−0.033	0.033
FI	−0.004	0.034	−0.066	0.066	−0.012	0.050	−0.099	0.099	−0.040	0.018	−0.036	0.036
FN	−0.021	0.003	−0.006	0.006	0.025	0.005	−0.011	0.011	−0.027	0.012	−0.023	0.023
GI	0.143	0.033	−0.065	0.065	0.066	0.021	−0.042	0.042	0.075	0.029	−0.058	0.058
GN	-0.029	0.012	−0.024	0.024	−0.014	0.021	−0.041	0.041	−0.057	0.015	−0.029	0.029
HI	−0.038	0.036	−0.070	0.070	−0.044	0.060	−0.117	0.117	−0.075	0.017	−0.034	0.034
HN	0.020	0.010	−0.020	0.020	0.054	0.013	−0.025	0.025	0.000	0.011	−0.022	0.022
II	0.103	0.029	−0.058	0.058	0.067	0.021	−0.040	0.040	0.053	0.023	−0.044	0.044
IN	−0.008	0.004	−0.008	0.008	0.032	0.007	−0.013	0.013	−0.019	0.012	−0.024	0.024
JI	−0.033	0.025	−0.048	0.048	−0.075	0.039	−0.076	0.076	−0.084	0.023	−0.045	0.045
JN	−0.002	0.005	−0.009	0.009	0.038	0.007	−0.014	0.014	−0.015	0.012	−0.023	0.023
KI	0.015	0.014	−0.027	0.027	−0.085	0.019	−0.038	0.038	−0.055	0.035	−0.068	0.068
KN	−0.015	0.006	−0.011	0.011	0.020	0.009	−0.018	0.018	−0.030	0.012	−0.024	0.024
LI	0.019	0.027	−0.053	0.053	0.008	0.035	−0.069	0.069	−0.019	0.019	−0.037	0.037
LN	−0.032	0.004	−0.008	0.008	0.008	0.008	−0.015	0.015	−0.043	0.012	−0.024	0.024
MI	0.068	0.054	−0.106	0.106	0.028	0.061	−0.120	0.120	0.017	0.023	−0.045	0.045
MN	0.000	0.014	−0.027	0.027	0.015	0.020	−0.039	0.039	−0.028	0.015	−0.029	0.029
NI	−0.111	0.014	−0.028	0.028	−0.195	0.023	−0.045	0.045	−0.176	0.033	−0.065	0.065
NN	0.011	0.043	−0.084	0.084	0.003	0.061	−0.120	0.120	−0.027	0.019	−0.038	0.038
OI	0.046	0.034	−0.067	0.067	0.032	0.039	−0.076	0.076	0.007	0.019	−0.038	0.038
ON	0.017	0.010	−0.020	0.020	0.048	0.013	−0.026	0.026	−0.004	0.012	−0.024	0.024
PI	0.077	0.028	−0.056	0.056	0.065	0.022	−0.043	0.043	0.039	0.019	−0.037	0.037
PN	0.010	0.005	−0.010	0.010	0.047	0.007	−0.014	0.014	−0.004	0.012	−0.023	0.023
QI	0.062	0.021	−0.041	0.041	0.012	0.024	−0.047	0.047	0.008	0.025	−0.049	0.049
QN	−0.006	0.007	−0.013	0.013	0.037	0.011	−0.021	0.021	−0.020	0.010	−0.019	0.019
RI	0.021	0.044	−0.087	0.087	−0.014	0.059	−0.116	0.116	−0.026	0.024	−0.046	0.046
RN	0.014	0.009	−0.018	0.018	0.035	0.012	−0.023	0.023	−0.009	0.015	−0.029	0.029

[a]LFDR: 95% lower false discovery rate.

[b]UFDR: 95% upper false discovery rate.

In the next section, we presented funnel plots constructed using the 3 methods for assessing centers' performance in providing services close to home for patients with end-stage kidney disease. The focus is to highlight how the variance estimation methods provide somewhat variable plots and how they affect interpretation and decision-making on the performance of centers in service provision.

Centers' Performance Using Funnel Plots

A summary funnel plot using the 3 methods is displayed in Figures 2-4. Each funnel plot has different variance estimates for the same underlying data. The funnel plots evaluate center-level performance in treating patients with end-stage kidney disease close to home by comparing the Log-SIR across different centers stratified by Indigenous status. The x-axis represents effective sample size (defined as a measure of the variability of the Log-SIRs for each center relative to the total variability of all Log-SMRs [28,28]), while the y-axis measures Log-SIR, indicating whether observed rates of receiving treatment close to home are higher or lower than expected. Centers within the upper and lower FDRs indicate expected performance in treating patients close to home (are in the region of average performance). The dashed lines forming funnels around the horizontal solid line (Log-SIR=0) indicate expected variation, with centers falling outside these limits exhibiting statistically significant differences from the norm.

Figure 2. Funnel plot using the delta method. Log-SIR: logarithm of the standardized incidence ratio.

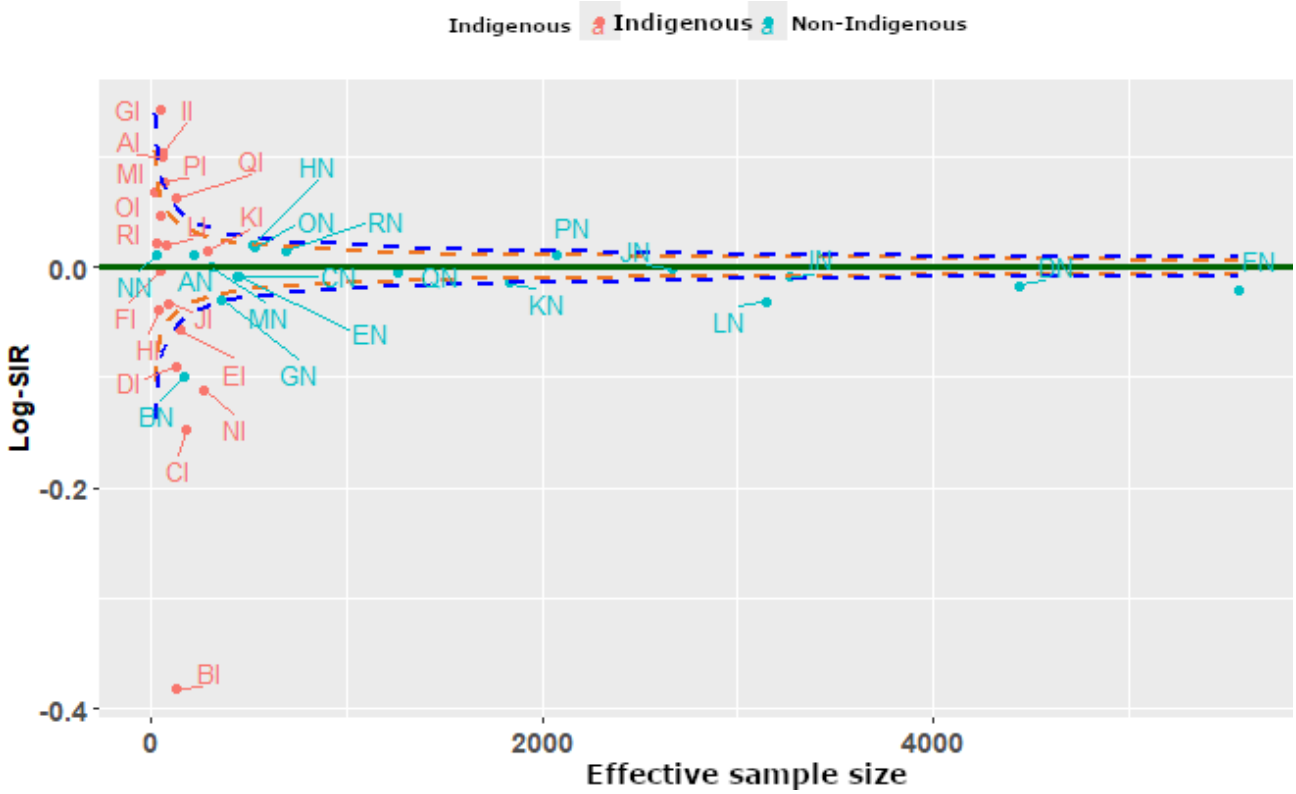


Figure 3. Funnel plot using the bootstrapping method. Log-SIR: logarithm of the standardized incidence ratio.

Figure 4. Funnel plot using the Bayesian Markov chain Monte Carlo method. Log-SIR: logarithm of the standardized incidence ratio.

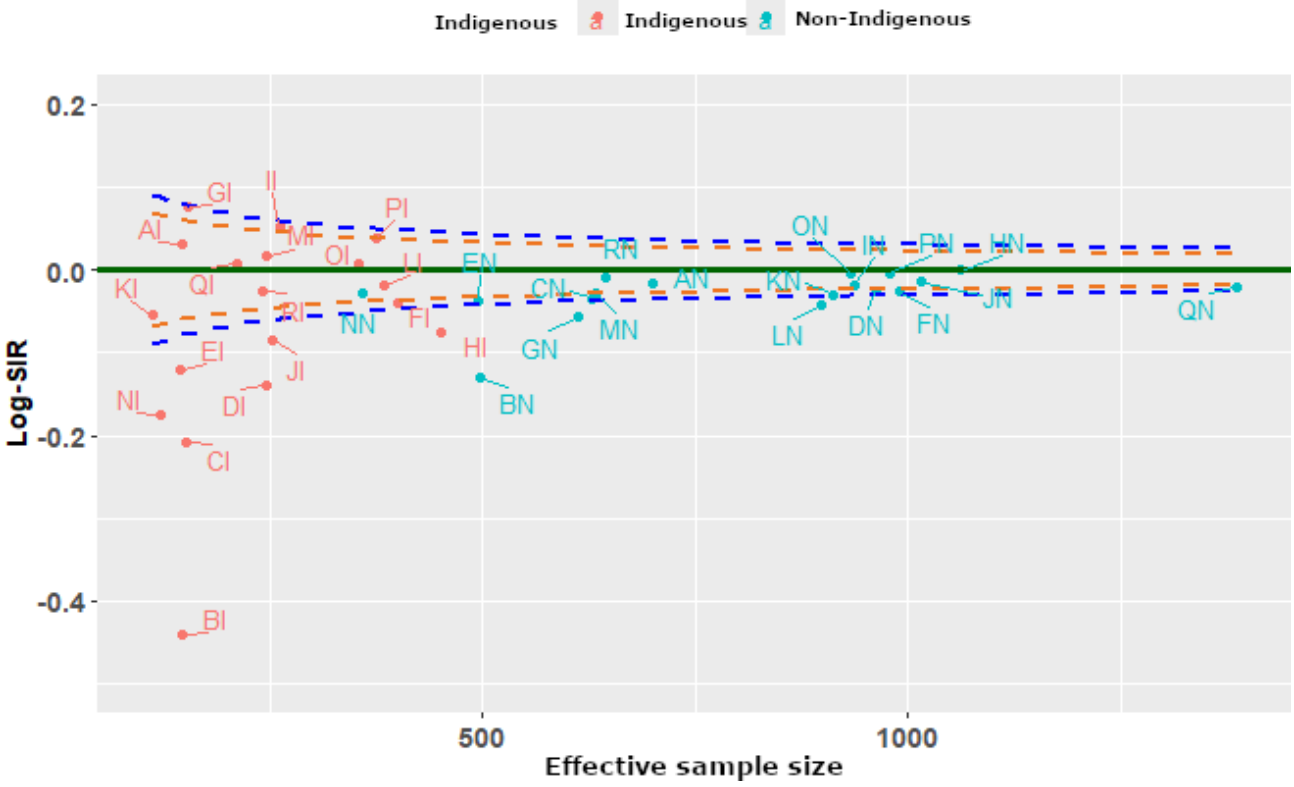


Figure 2 presents a funnel plot that compares the performance of centers using the delta method for estimating the variance of the Log-SIR. Centers within the upper and lower FDRs indicate expected performance in treating patients close to home (are in the region of average performance). The dashed lines forming funnels around the horizontal solid line (Log-SIR=0) indicate expected variation, with centers falling outside these limits exhibiting statistically significant differences from the norm.

Using this approach, 6 centers, BI, BN, CI, DI, EI, and NI, were low performing, while 3 centers, GI, II, and QI, were higher-than-average performers. The remaining centers lie

within the FDRs, being average performers in treating patients close to home. Center BI shows the lowest Log-SIR, suggesting exceptionally lower performance in treating patients close to home. Overall, larger centers exhibit more stable Log-SIR values, while smaller centers experience greater variation, reinforcing the importance of center size in the assessment of centers' performance in treating patients close to home. Using this method, the variance of Log-SIR appears relatively low, with most values concentrated around zero. Some extreme values (outliers) are present on the left-hand side, indicating a few centers with more deviation. The spread of points suggests that this method results in a tighter distribution of Log-SIR.

Figure 3 compares centers using the bootstrapping approach. Using this method, 7 centers, BI, BN, CI, DI, EI, NI, and KI, were lower-than-average performers. However, 12 centers, GI, II, PI, DN, FN, HN, IN, JN, ON, PN, QN, and RN, were found to be higher-than-average performing. The remaining centers lie within the FDRs, being average performers in treating patients close to home. Notably, BI remains an outlier with the lowest Log-SIR, reflecting exceptionally low performance in treating patients close to home. Using bootstrap, the variance is slightly larger compared with the first plot. The spread of Log-SIR values is more noticeable, with a wider range of deviations from zero. More centers have larger deviations, particularly on the left side, compared with the delta method.

Figure 4 presents a funnel plot that compares the performance of centers using the Bayesian approach for estimating the variance of the Log-SIR. Accordingly, 11 centers, BI, BN, CI, DI, EI, GN, HI, JI, KI, NI, and LN, were found low-than-expected performers, and no center was found to be top performing in treating patients close to home. Larger centers exhibit more stable Log-SIR values, reinforcing the reliability of their performance assessments. Using the Bayesian approach, the variance of Log-SIR is still larger than the first plot but somewhat comparable with the second. The spread is not as extreme as in the second plot, but it still shows noticeable deviations. There are clear differences in the spread of values across regions.

The delta method results in the least variance in Log-SIR, while the bootstrapping method has the highest variance, with a wider spread of values. Clearly, the Bayesian approach has an intermediate variance, showing more spread than the first method but less than the second.

Discussion

Overview

Our study results highlight center-level differences in treating patients close to home, and this is coupled with variability in variance estimation by the 3 methods. The stability of the Log-SIR using the Bayesian approach may be due to the method borrowing strength from prior beliefs, which are summarized using probability distributions that smooth variability in estimation.

In health care providers' performance assessment, standardized incidence ratios (SIRs) and standardized mortality ratios (SMRs) are essential tools used to assess whether observed rates of disease or death deviate from what is expected. Accurate estimation of variance in these ratios is crucial as it affects decision-making regarding providers' performance, resource allocation, and quality improvement strategies. In this study, we compared 3 methods, namely, the delta method, bootstrapping, and Bayesian approach, to estimate the variance of the Log-SIR given by equation 3 and considered funnel plot approaches to build FDRs around the Log-SIR using these 3 variance estimators. The variance estimation methods have been widely discussed in statistical literature. Gelman et al [43] emphasize that Bayesian methods, particularly MCMC, provide more stable estimates due to their ability to incorporate prior information and reduce uncertainty. Similarly, Efron and Tibshirani [11] discuss bootstrapping as a flexible but sometimes overly variable approach, which aligns with our findings of increased variance in bootstrapped estimates.

The delta method is frequently used in epidemiology for variance estimation [44]. It provides an efficient and straightforward way of estimating the variance of Log-SIR or Log-SMR, especially when the distribution of the underlying data was correctly specified. This method can be computationally efficient, but its accuracy may suffer in cases where the underlying distribution deviates significantly from the assumed form [45]. When applied in health care decision-making, such as assessing the performance of hospitals based on SMRs, the delta method may underestimate variance if assumptions are violated. This could lead to incorrect conclusions regarding the performance of health care providers.

Variance estimation using the delta method for metrics other than SMR has been used intensively. For instance, Normand and Shahian [46] applied the delta method to approximate the variance of demographic parameters in avian biology studies. Although not directly related to health care, this study illustrates the broader applicability of the delta method in estimating variances of complex ratios. Also, Lee et al [47] compared the Green, delta, and Monte Carlo methods for calculating the 95% CI for population-attributable fraction. In addition, Sauer et al [48] applied the delta method for variance estimation for effective coverage measures. There is limited study that directly applied the delta method in the estimation of Log-SIR used in the assessing performance of health care providers in the provision of health services for a given outcome.

Bootstrapping, on the contrary, has the advantage of not relying on distributional assumptions and can be used to directly estimate the distribution of Log-SIR or Log-SMR. This can lead to more robust variance estimates, particularly in settings with small sample sizes or unknown distributions. By resampling, bootstrapping accounts for sampling variability and can help improve the precision of performance assessments [11]. For instance, Kasza et al [28] used bootstrapping for evaluating the performance of Australian and New Zealand intensive care units in 2009 and 2010 quantified by the standardized mortality ratio. Moreover, Walters and Campbell [49] used bootstrap methods for analyzing health-related quality-of-life outcomes used in clinical trials as primary outcome measures. They found that certain bootstrap methods provided more accurate variance estimates, especially when the distribution of the outcome is unknown or ordinal scale.

By contrast, Bayesian methods provide a full posterior distribution for variance estimates, allowing for the incorporation of prior knowledge, such as expert opinion or historical data on hospital performance. This can lead to more flexible and informative variance estimation, especially when data are sparse or prior knowledge is available. Bayesian methods can also be used to model hierarchical structures (eg, hospitals within regions), providing more precise estimates of performance at various levels [32].

A study by George et al [50] applied Bayesian hierarchical models to estimate hospital performance in the Hospital Compare model for acute myocardial infarction mortality. They found that indirect standardization fails to adequately control for differences in patient risk factors and systematically underestimates mortality rates at the low-volume hospitals.

Below, we have summarized the variability in variance estimates and their implications on funnel plots and epidemiological studies.

Variability in Log SIR Variance Estimates

The 3 methods yield different variance estimates for the same underlying data. Bootstrapping tends to produce higher variance estimates due to the nature of resampling, which can exaggerate variability, particularly in small samples [51]. By contrast, Bayesian (MCMC) estimates tend to be more stable, benefiting from prior distributions that help regularize estimates, a characteristic also observed in Bayesian hierarchical models for disease mapping [52]. The delta method, being a first-order approximation, is the most conservative, often producing the lowest variance estimates, which may lead to underestimation in complex data structures [32]. These differences highlight the importance of choosing an estimation method suited to the underlying data characteristics and sample size.

In our study, the variance estimates differ across methods, with bootstrapping tending to show more extreme values (both high and low) compared to the other 2 methods. MCMC appears to provide more stable and generally lower variance estimates compared to bootstrapping. The delta method is relatively consistent but tends to lie between the MCMC and bootstrap estimates. Some centers have noticeably higher variance estimates for all 3 methods (eg, locations where green dots are well above the others). This suggests that uncertainty in Log-SIR estimation varies by center, possibly due to differences in sample size, population characteristics, or underlying risk factors. Bootstrapping shows more variability, which is expected since it resamples data and may amplify variability in small samples. MCMC provides more stable estimates, benefiting from Bayesian shrinkage and prior information incorporation. The delta method is computationally efficient but may underestimate variance in some cases (eg, when normality assumptions are violated) [53]. Centers with higher variance estimates (especially under bootstrapping) suggest that Log-SIR estimates are more uncertain there, which should be considered when making public health decisions. If variance estimates are too high, it may indicate the need for larger sample sizes or improved data collection in those centers.

Impact on Funnel Plots

The funnel plots illustrate how these methods influence the distribution of Log-SIR estimates. The Bayesian approach exhibits a more stabilized pattern, particularly at smaller sample sizes, where shrinkage effects help reduce extreme values. This aligns with findings from Spiegelhalter et al [54], who demonstrated that Bayesian hierarchical modeling effectively mitigates overdispersion in epidemiological data. Conversely, the bootstrapping approach results in greater spread at smaller sample sizes, reflecting its sensitivity to sample fluctuations. Similar findings have been reported in comparative studies on variance estimation methods, where bootstrapping is noted to introduce greater variability but remains valuable for robust uncertainty estimation [11]. Although both methods show convergence of Log-SIR estimates toward zero as sample sizes increase, bootstrapping maintains slightly higher variance, reinforcing the need for careful interpretation in small-sample studies.

Implications for Epidemiological Studies

The choice of variance estimation method has significant implications for epidemiological research. Bayesian methods offer improved stability and are particularly useful when incorporating prior knowledge is beneficial. Studies have shown that Bayesian approaches reduce estimation bias and enhance interpretability in spatial epidemiology [55]. Bootstrapping, despite its higher variability, remains a valuable tool for robust uncertainty estimation, especially when parametric assumptions may not hold [56]. Meanwhile, the delta method, though computationally simple, may underestimate variance, making it less reliable for complex data scenarios, as previously noted in statistical inference literature [32]. These findings align with broader discussions on variance estimation in epidemiology, emphasizing the trade-offs between robustness, computational efficiency, and precision [57].

Principal Findings

These findings highlight the importance of selecting an appropriate variance estimation method depending on the study context. Bayesian methods may be preferable when stability and regularization are critical, while bootstrapping is useful for assessing variability in more flexible settings. The delta method should be used cautiously, particularly when dealing with skewed or complex distributions. Future research should explore hybrid approaches that combine the strengths of these methods for more robust inference [11,32].

Our results showed that Bayesian approaches provided more conservative estimates with tighter credible intervals, particularly in hospitals with small case volumes. We demonstrated that Bayesian MCMC outperforms the other methods in terms of lower variance and MSE, making it the preferred choice for estimating Log-SIR variance when computational resources permit.

Limitations

Our study has several limitations. First, while understanding the differences between variance estimation methods is crucial for assessing the reliability of SIR estimates across different centers, we did not consider how model choice influences

variance estimates and hence the resulting statistical inference. That is, we only used hierarchical logistic regression model for modeling the binary individual-level outcome. Therefore, we did not explore the implication of using the Poisson model for aggregated data on the resulting variance estimates using the 3 methods. Second, we considered only nonparametric bootstrapping, and the implications of parametric bootstrapping were not assessed. Third, we did not consider other transformations than logarithmic transformations and their effects on the interpretation of providers' performance. For instance, Quaresma et al [12] investigated the implications of identity(log), complementary log-log, logit, and logarithmic transformation in their study of cancer survival. Finally, within the random effects logistic regression, we considered only logit link, and other links such as probit and complementary log-log link were not considered here.

Conclusions

In conclusion, the choice of variance estimation method plays a significant role in how health care providers' performance is assessed. While each method has its strengths and weaknesses, bootstrapping and Bayesian approaches generally provide more reliable estimates of uncertainty compared to the delta method. However, the choice of method should consider computational resources, data structure, and the available prior knowledge for Bayesian methods. Decision-makers should be aware of the implications of variance estimation on conclusions regarding provider performance, which can influence policy, resource allocation, and quality improvement initiatives in health care settings. In terms of decision-making, the choice of variance estimation method can affect the conclusions drawn about the performance of health care providers. Using the delta method may lead to an underestimation of uncertainty, especially when the data do not meet distributional assumptions. Bootstrapping, while more robust, may be computationally intensive, especially with large datasets. Bayesian methods, with their flexibility and ability to incorporate prior knowledge, can be powerful tools but require careful specification of priors and may be computationally demanding.

Acknowledgments

This study is funded by the National Health and Medical Research Council of Australia (GNT1158075). We are grateful to the Australian National Health and Medical Research Council (NHMRC) for supporting the "Return to Country" project (GNT1158075), which this methodological paper is a part of. The data reported here have been supplied by the Australia and New Zealand Dialysis and Transplant Registry (ANZDATA). The interpretation and reporting of these data are the responsibility of the authors and in no way should be seen as an official policy or interpretation of the Australia and New Zealand Dialysis and Transplant Registry.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Model and logarithm of standardized incidence ratio definitions.
[DOCX File, 17 KB - xmed_v6i1e77415_app1.docx]

Multimedia Appendix 2
Delta method.
[DOCX File, 28 KB - xmed_v6i1e77415_app2.docx]

Multimedia Appendix 3
Bayesian approach.
[DOCX File, 23 KB - xmed_v6i1e77415_app3.docx]

References

1. National health reform: progress and delivery. : Australian Department of Health and Ageing; 2011 URL: https://catalogue.nla.gov.au/catalog/5816021 [accessed 2025-09-30]
2. Australian Government. National Health Reform Agreement (NHRA) – Long-term Health Reforms – Roadmap.: Commonwealth of Australia; 2021. URL: https://www.health.gov.au/sites/default/files/documents/2021/10/national-health-reform-agreement-nhra-long-term-health-reforms-roadmap_0.pdf [accessed 2025-09-23]
3. Spiegelhalter DJ. Funnel plots for comparing institutional performance. Stat Med 2005 Apr 30;24(8):1185-1202. [doi: 10.1002/sim.1970] [Medline: 15568194]
4. Goldstein H, Spiegelhalter DJ. Statistical aspects of institutional performance: league tables and their limitations (with discussion). Journal of the Royal Statistical Society; Series A 1996;159:385-444. [doi: 10.2307/2983325]
5. Shewhart WA. The application of statistics as an aid in maintaining quality of a manufactured product. J Am Stat Assoc 1925 Dec;20(152):546-548. [doi: 10.1080/01621459.1925.10502930]

6. McDonald SP. Australia and New Zealand dialysis and transplant registry. Kidney Int Suppl (2011) 2015 Jun;5(1):39-44. [doi: 10.1038/kisup.2015.8] [Medline: 26097784]
7. Verburg IW, Holman R, Peek N, Abu-Hanna A, de Keizer NF. Guidelines on constructing funnel plots for quality indicators: a case study on mortality in intensive care unit patients. Stat Methods Med Res 2018 Nov;27(11):3350-3366. [doi: 10.1177/0962280217700169] [Medline: 28330409]
8. Quaresma M, Coleman MP, Rachet B. Funnel plots for population-based cancer survival: principles, methods and applications. Stat Med 2014 Mar 15;33(6):1070-1080. [doi: 10.1002/sim.5953] [Medline: 24038332]
9. Vasilevskis EE, Kuzniewicz MW, Dean ML, et al. Relationship between discharge practices and intensive care unit in-hospital mortality performance: evidence of a discharge bias. Med Care 2009 Jul;47(7):803-812. [doi: 10.1097/MLR.0b013e3181a39454] [Medline: 19536006]
10. Mazzucco W, Cusimano R, Zarcone M, Mazzola S, Vitale F. Funnel plots and choropleth maps in cancer risk communication: a comparison of tools for disseminating population-based incidence data to stakeholders. BMJ Open 2017 Mar 30;7(3):e011502. [doi: 10.1136/bmjopen-2016-011502] [Medline: 28363917]
11. Efron B, Tibshirani RJ. An Introduction to the Bootstrap: Chapman & Hall/CRC; 1993.
12. Quaresma M, Coleman MP, Rachet B. Funnel plots for population-based cancer survival: principles, methods and applications. Statist Med 2014 Mar 15;33(6):1070-1080. [doi: 10.1002/sim.5953] [Medline: 24038332]
13. Powell LA. Approximating variance of demographic parameters using the delta method: a reference for avian biologists. Condor 2007 Nov 1;109(4):949-954. [doi: 10.1093/condor/109.4.949]
14. Hosmer DW, Lemeshow S. Confidence interval estimates of an index of quality performance based on logistic regression models. Stat Med 1995 Oct 15;14(19):2161-2172. [doi: 10.1002/sim.4780141909] [Medline: 8552894]
15. Austin PC. The failure of four bootstrap procedures for estimating confidence intervals for predicted-to-expected ratios for hospital profiling. BMC Med Res Methodol 2022 Oct 14;22(1):271. [doi: 10.1186/s12874-022-01739-x] [Medline: 36241973]
16. Ventrucci M, Scott EM, Cocchi D. Multiple testing on standardized mortality ratios: a Bayesian hierarchical model for FDR estimation. Biostatistics 2011 Jan;12(1):51-67. [doi: 10.1093/biostatistics/kxq040] [Medline: 20577014]
17. Sukul D, Seth M, Thompson MP, et al. Hospital and operator variation in cardiac rehabilitation referral and participation after percutaneous coronary intervention: insights from blue cross blue shield of Michigan cardiovascular consortium. Circ Cardiovasc Qual Outcomes 2021 Nov;14(11):e008242. [doi: 10.1161/CIRCOUTCOMES.121.008242] [Medline: 34749515]
18. Devitt J, McMasters A. Living on Medicine: A Cultural Study of End-Stage Renal Disease Among Aboriginal People: IAD Press; 1998.
19. Anderson K, Cunningham J, Devitt J, et al. "Looking back to my family": Indigenous Australian patients' experience of hemodialysis. BMC Nephrol 2012;13:114. [doi: 10.1186/14712369-13-114]
20. Hughes JT, Dembski L, Kerrigan V, Majoni SW, Lawton PD, Cass A. Gathering perspectives - finding solutions for chronic and end stage kidney disease. Nephrology (Carlton) 2018 Feb;23 Suppl 1:5-13. [doi: 10.1111/nep.13233] [Medline: 29436104]
21. Marley JV, Dent HK, Wearne M, et al. Haemodialysis outcomes of Aboriginal and Torres Strait Islander patients of remote Kimberley region origin. Med J Aust 2010 Nov 1;193(9):516-520. [doi: 10.5694/j.1326-5377.2010.tb04035.x] [Medline: 21034385]
22. ANZDATA Registry. 38th Report, Chapter 12: Indigenous People and End Stage Kidney Disease. 2016 URL: https://www.anzdata.org.au/wp-content/uploads/2023/10/c12_anzdata_indigenous_v3.0_201600128_web.pdf [accessed 2025-09-30]
23. Jones Y, Truong M, Preece C, et al. Study protocol: Return to Country, an Australia-wide prospective observational study about returning First Nations renal patients home. BMJ Open 2024 Nov 24;14(11):e095727. [doi: 10.1136/bmjopen-2024-095727] [Medline: 39581708]
24. Greenland S. Modeling and variable selection in epidemiologic analysis. Am J Public Health 1989 Mar;79(3):340-349. [doi: 10.2105/ajph.79.3.340] [Medline: 2916724]
25. Goldstein H, Spiegelhalter DJ. League tables and their limitations: statistical issues in comparisons of institutional performance. J R Stat Soc Ser A Stat Soc 1996;159(3):385-443. [doi: 10.2307/2983325]
26. Gelman A, Hill J. Data Analysis Using Regression and Multilevel/Hierarchical Models: Cambridge University Press; 2007.
27. Lawson AB. Bayesian Disease Mapping: Hierarchical Modeling in Spatial Epidemiology, 3rd edition: CRC Press; 2018.
28. Kasza J, Moran JL, Solomon PJ, ANZICS-Australian New Zealand Intensive Care Society Centre for Outcome and Resource Evaluation-CORE. Evaluating the performance of Australian and New Zealand intensive care units in 2009 and 2010. Stat Med 2013 Sep 20;32(21):3720-3736. [doi: 10.1002/sim.5779] [Medline: 23526209]
29. Normand SLT, Shahian DM, Krumholz HM. Statistical and clinical aspects of hospital outcomes profiling. Statist Sci 2007;22(2):206-226. [doi: 10.1214/088342307000000096]
30. Clark DE, Moore L. Multilevel modeling. In: Li G, Baker S, editors. Injury Research: Springer; 2011. [doi: 10.1007/978-1-4614-1599-2_23]
31. Yang X, Peng B, Chen R, et al. Statistical profiling methods with hierarchical logistic regression for healthcare providers with binary outcomes. J Appl Stat 2014;41(1):46-59. [doi: 10.1080/02664763]
32. Casella G, Berger RL. Statistical Inference, 2nd edition 2002.
33. Bickel PJ, Doksum KA. Mathematical Statistics: Basic Ideas and Selected Topics, 2nd edition: Pearson; 2006, Vol. 1.

34. Vaart AW. Asymptotic Statistics: Cambridge University Press; 2000.
35. Boos DD, Stefanski LA. Essential Statistical Inference: Theory and Methods: Springer; 2013.
36. R Core Team. R: a language and environment for statistical computing. R Foundation for Statistical Computing. 2023. URL: https://www.R-project.org/ [accessed 2025-09-23]
37. Bates D, Maechler M, Bolker B, Walker S. Lme4: linear mixed-effects models using 'Eigen' and S4. R package version 11-34. 2023. URL: https://cran.r-project.org/web/packages/lme4/index.html [accessed 2025-09-23]
38. Canty A, Ripley B. Boot: bootstrap functions (originally by Angelo Canty for S). R package version 13-30. 2024. URL: https://CRAN.R-project.org/package=boot [accessed 2025-09-23]
39. Wickham H. Ggplot2: elegant graphics for data analysis. R package version 351.: Springer; 2016. URL: https://CRAN.R-project.org/package=ggplot2 [accessed 2025-09-23]
40. Slowikowski K. Ggrepel: automatically position non-overlapping text labels with ggplot2. R package version 095. 2024. URL: https://CRAN.R-project.org/package=ggrepel [accessed 2025-09-23]
41. R Core Team. Parallel: support for parallel computation in R. R package included in base R, version 440. 2024. URL: https://stat.ethz.ch/R-manual/R-devel/library/parallel/doc/parallel.pdf [accessed 2025-09-23]
42. Lüdecke D. SjPlot: data visualization for statistics in social science. R package version 2815. 2023. URL: https://CRAN.R-project.org/package=sjPlot [accessed 2025-09-23]
43. Gelman A, Carlin JB, Stern HS, Dunson DB, Vehtari A, Rubin DB. Bayesian Data Analysis, 3rd edition: Chapman and Hall/CRC; 2013. [doi: 10.1201/b16018]
44. Rothman KJ, Greenland S, Lash TL. Modern Epidemiology, 3rd edition: Lippincott Williams & Wilkins; 2008.
45. Corlu CG, Akcay A, Xie W. Stochastic simulation under input uncertainty: a review. Operations Research Perspectives 2020;7:100162. [doi: 10.1016/j.orp.2020.100162]
46. Normand SLT, Shahian DM. Statistical and clinical aspects of hospital outcomes profiling. Statist Sci 2007 May;22(2):206-226. [doi: 10.1214/088342307000000096]
47. Lee S, Moon S, Kim K, et al. A comparison of Green, delta, and Monte Carlo methods to select an optimal approach for calculating the 95% confidence interval of the population-attributable fraction: guidance for epidemiological research. J Prev Med Public Health 2024 Sep;57(5):499-507. [doi: 10.3961/jpmph.24.272] [Medline: 39265631]
48. Sauer SM, Pullum T, Wang W, Mallick L, Leslie HH. Variance estimation for effective coverage measures: a simulation study. J Glob Health 2020 Jun;10(1):010506. [doi: 10.7189/jogh.10.010506] [Medline: 32257160]
49. Walters SJ, Campbell MJ. The use of bootstrap methods for analysing health-related quality of life outcomes (particularly the SF-36). Health Qual Life Outcomes 2004 Dec 9;2:70. [doi: 10.1186/1477-7525-2-70] [Medline: 15588308]
50. George EI, Ročková V, Rosenbaum PR, Satopää VA, Silber JH. Mortality rate estimation and standardization for public reporting: Medicare's Hospital Compare. J Am Stat Assoc 2017 Jul 3;112(519):933-947. [doi: 10.1080/01621459.2016.1276021]
51. Davison AC, Hinkley DV. Bootstrap Methods and Their Application: Cambridge University Press; 1997. URL: https://www.cambridge.org/core/books/bootstrap-methods-and-their-application/ED2FD043579F27952363566DC09CBD6A
52. Clayton D, Kaldor J. Empirical Bayes estimates of age-standardized relative risks for use in disease mapping. Biometrics 1987 Sep;43(3):671-681. [Medline: 3663823]
53. Gupta RS, Carrión-Carire V, Weiss KB. The widening black/white gap in asthma hospitalizations and mortality. J Allergy Clin Immunol 2006 Feb;117(2):351-358. [doi: 10.1016/j.jaci.2005.11.047] [Medline: 16461136]
54. Spiegelhalter DJ, Best NG, Carlin BP, Van Der Linde A. Bayesian measures of model complexity and fit. J R Stat Soc Ser B Methodol 2002 Oct 1;64(4):583-639. [doi: 10.1111/1467-9868.00353]
55. Best N, Richardson S, Thomson A. A comparison of Bayesian spatial models for disease mapping. Stat Methods Med Res 2005 Feb;14(1):35-59. [doi: 10.1191/0962280205sm388oa] [Medline: 15690999]
56. Carpenter J, Bithell J. Bootstrap confidence intervals: when, which, what? A practical guide for medical statisticians. Stat Med 2000 May 15;19(9):1141-1164. [doi: 10.1002/(sici)1097-0258(20000515)19:9<1141::aid-sim479>3.0.co;2-f] [Medline: 10797513]
57. Gustafson P. Measurement Error and Misclassification in Statistics and Epidemiology: Impacts and Bayesian Adjustments: CRC Press; 2003. URL: https://www.taylorfrancis.com/books/mono/10.1201/9780203502761/measurement-error-misclassification-statistics-epidemiology-paul-gustafson [accessed 2025-09-23]

Abbreviations

ANZDATA: Australia and New Zealand Dialysis and Transplant Registry
FDR: false discovery rate
HREC: Human Research Ethics Committee
Log: logarithm
Log-SIR: logarithm of the standardized incidence ratio
MCMC: Markov chain Monte Carlo
MSE: mean squared error

SIR: standardized incidence ratio
SMR: standardized mortality ratio

Edited by S Tungjitviboonkun; submitted 13.05.25; peer-reviewed by E Oluwagbade; revised version received 11.08.25; accepted 30.08.25; published 09.10.25.

Please cite as:
Woldeyohannes S, Jones Y, Lawton P
Estimating Variance of Log Standardized Incidence Ratios Assessing Health Care Providers' Performance: Comparative Analysis Using Bayesian, Bootstrap, and Delta Method Approaches
JMIRx Med 2025;6:e77415
URL: https://xmed.jmir.org/2025/1/e77415
doi:10.2196/77415

Evaluating the Financial Factors Influencing Maternal, Newborn, and Child Health in Africa: Tobit Regression and Data Envelopment Analysis

Youssef Er-Rays[1]; Meriem M'dioud[2]; Hamid Ait-Lemqeddem[2]; Badreddine El Moutaqi[1]

[1]Polydisciplinary Faculty of Larache, Abdelmalek Essaadi University, Tetouan, Morocco

[2]École nationale des sciences appliquées, Ibn Tofail University, Kenitra, Morocco

Corresponding Author:
Youssef Er-Rays
Polydisciplinary Faculty of Larache, Abdelmalek Essaadi University, Tetouan, Morocco

Related Articles:

Companion article: https://www.medrxiv.org/content/10.1101/2024.02.22.24303217v1

Companion article: https://med.jmirx.org/2025/1/e85382

Companion article: https://med.jmirx.org/2025/1/e85383

Companion article: https://med.jmirx.org/2025/1/e85578

Abstract

Background: Despite international efforts, maternal, newborn, and child health (MNCH) outcomes in Africa continue to lag due to inefficient health systems and underperforming financial frameworks. Financial factors—such as total health expenditure, health coverage indices, and spending per capita—are key but understudied drivers of MNCH service efficiency.

Objective: This study investigates the extent to which financial inputs influence the technical efficiency of MNCH service delivery across 46 African countries. The aim is to generate evidence for health financing policies that can enhance both efficiency and health equity.

Methods: We adopted a 2-stage analytical framework. First, data envelopment analysis using a variable returns-to-scale, input-oriented model was applied to measure technical efficiency. Second, Tobit regression identified the financial determinants of inefficiency. Explanatory variables included current health expenditures, a health coverage index, and current health expenditures per capita.

Results: Only 12 of 46 countries (26%) achieved full technical efficiency (efficiency score=1), while the rest (n=34, 74%) were inefficient, with a mean score of 0.849. Efficiency was notably lower in low-income countries (mean 0.810) compared to upper-middle-income countries (mean 0.940). Tobit regression showed that increased current health expenditure significantly reduced inefficiency (β=–.0811; *P*=.001). Conversely, a higher health coverage index unexpectedly increased inefficiency (β=.0155; *P*=.001), suggesting that expanded coverage without improved governance or resource capacity may strain systems. Health expenditure per capita was not statistically significant. Model 2 demonstrated stronger explanatory power (pseudo R^2=0.8943).

Conclusions: Financial factors, particularly total health expenditure, play a decisive role in shaping MNCH efficiency across African nations. However, expanding health coverage without parallel improvements in system governance may exacerbate inefficiencies. To enhance MNCH outcomes, policy efforts must focus on increasing and strategically allocating financial resources while strengthening institutional accountability and performance.

(JMIRx Med 2025;6:e59703) doi:10.2196/59703

KEYWORDS

financial determinants; maternal, newborn, and child health; health care efficiency; Africa; health expenditure; data envelopment analysis; Tobit regression

Introduction

Maternal, newborn, and child health (MNCH) is a crucial aspect of global well-being, as outlined in Sustainable Development Goals (SDGs) 3.1 and 3.2. Despite global efforts, Africa continues to face the highest MNCH mortality rates, with 287,000 women dying in 2020 and 5 million children dying in 2021. The World Health Organization (WHO) highlights significant regional imbalances in MNCH outcomes across Africa [1]. A WHO report predicts a slowdown in Africa's progress against maternal and infant mortality over the past decade [2]. The Atlas of African Health Statistics 2022 reveals that increased investment is needed to accelerate progress toward the SDG on health. Maternal mortality is one of the most difficult targets to achieve, with an estimated 390 women dying in childbirth for every 100,000 live births by 2030 [2]. To reach the SDG target, Africa needs an 86% reduction from 2017 rates, which is unrealistic at the current rate of decline. The region's infant mortality rate stands at 72 per 1000 live births, with an expected 54 deaths per 1000 live births by 2030 [2]. Although Africa has made significant progress in some areas, such as vaccine coverage, the slowdown has been exacerbated by the COVID-19 pandemic disrupting crucial health services and a resurgence in vaccine-preventable disease outbreaks. Inadequate investment in health and funding for health programs are major drawbacks to meeting the SDG on health. These persistent challenges are exacerbated by weak leadership, corruption, and systemic weaknesses within African health systems, leading to profound inequalities despite overall global mortality halving by 2021 [3-5].

Addressing this critical MNCH challenge requires a deep understanding of how health care resources are used. Therefore, investigating the efficiency of health care systems is paramount to identifying areas where resource allocation can be optimized to improve health outcomes, especially in contexts with limited resources, like many African countries. Global health care system evaluations are essential for pinpointing such improvement areas, and recent research consistently focuses on health care performance [6-11].

Data envelopment analysis (DEA) models are widely recognized and used as tools for evaluating health care efficiency. These models effectively assess the performance of decision-making units (DMUs) by measuring their ability to generate multiple outputs from multiple inputs [12,13]. Research using DEA to compare health care systems has been extensive across various regions. For instance, studies have explored efficiency in 34 Organisation for Economic Co-operation and Development member nations [14], 30 European states [15], 20 Arab countries [16], the Middle East and North Africa region [17], 18 nations within the Middle East and North Africa region [18], and 46 Asian countries [19]. A recurring theme in these studies is the consistent use of variables such as health spending, the number of doctors, and the number of hospital beds as inputs, while life expectancy and infant mortality rates frequently serve as key outcome measures. This methodological consistency, despite the diversity of regions and specific approaches, significantly contributes to a comprehensive understanding of health care system efficiency on a global scale.

However, despite the global prevalence of DEA in health care efficiency studies, research specifically focusing on health system efficiency in Africa remains limited. Musoke et al [20] noted a relative drop in DEA research on the continent, although this approach has seen increased adoption in African studies over the last decade. This gap highlights a critical need for more region-specific analysis to understand and address the unique efficiency challenges within African health systems, particularly in the context of improving MNCH outcomes.

In 2023, Musoke et al [20] compared the health systems of 29 of the least developed African countries. The inputs included domestic general government health, domestic private health, external health, and out-of-pocket health. Outputs included the under-5 survival rate, maternal survival ratio, life expectancy at birth, and infant survival rate.

Top et al [21] examined 36 African health care systems, considering health expenditures in the gross domestic product; medical professionals, nurses, and bed capacity per 1000 individuals; the unemployment rate; and the Gini coefficient. Life expectancy at birth and 1/(infant mortality rate) were the study's output variables.

Two studies assessed the effectiveness of health care systems in 45 African countries using infant mortality rates and per capita health expenditure and real gross domestic product [22,23]. Another study found that health care infrastructure in sub-Saharan African countries is ineffective due to management weaknesses at multiple levels [24]. Kirigia et al [25,26] investigated efficiency using factors like per capita total health expenditure, adult literacy rate, and male and female life expectancies as outcome variables [25-27].

In a separate study, Arhin et al [28] assessed the ability of the health system to achieve the universal health coverage (UHC) goal by drawing evidence from 30 African countries. The study integrated per capita health spending and physician and hospital data as inputs, with the UHC index serving as the output metric.

However, Qu et al [29] undertook a comparative analysis encompassing 49 African countries from 2000 to 2017. They introduced an innovative methodology that amalgamates DEA with the Gini coefficient to assess the efficacy of technology inequality in addressing environmental issues.

Recent studies have evaluated the efficiency of maternal and child health services using DEA, particularly in developing and middle-income countries such as Morocco. One study, by Youssef et al, used a 2-stage DEA model on 76 MNCH primary health care units in Morocco, revealing an average efficiency score of 0.779 under constant returns to scale (CRS). The study also found significant disparities across provinces, with Boujdour ranking the lowest. Tobit regression revealed that rural health dispensaries and support programs for high-risk pregnancies positively influenced efficiency [9]. A longitudinal study by the same authors used a longitudinal dataset covering 9 years, applying both input- and output-oriented DEA, the Malmquist index, and Tobit regression to assess hospital performance. The results showed an average input-oriented efficiency score of 0.76 and an output-oriented score of 0.23, with mixed productivity trends [11].

Several studies have used DEA to assess MNCH efficiency. Other studies explored efficiency through alternative methods, such as a virtual reality tool that reduced pediatric magnetic resonance imaging anesthesia costs [30], a parental training program that shortened neonatal intensive care unit stays [31], and a mobile health app that lowered asthma hospitalization expenses. These scalable interventions optimized MNCH budgets by reducing resource consumption.

Comprehensive studies focusing specifically on MNCH across African countries—or even within individual nations—remain scarce, with the notable exception of work by Er-Rays and colleagues [32]. This underscores the originality of this paper, which conducts a novel analysis evaluating the financial determinants influencing MNCH in Africa through the application of DEA and Tobit regression.

The literature reviews an assessment of health care system efficiency in other regions, pointing out that it requires a careful selection of inputs, outputs, and explanatory variables. Most of the studies used inputs, which included health care expenditures, health care personnel (doctors, nurses, midwives), hospital beds, and health facilities. The frequently used outputs consisted of life expectancy, health care utilization, and health outcomes. The most used explanatory variables included financial factors, governance, geographic location, infrastructure, and technology. However, most of these studies neglected to consider the maternal mortality rate, stillbirth rate, neonatal mortality rate, and number of births attended by skilled health personnel. Hence, this original paper addresses the technical efficiency of MNCH in Africa.

Motivated by the imperative to achieve SDGs 3.1 and 3.2 by 2030, it is paramount to assess the effectiveness of health systems in Africa, emphasizing the critical need for Africans to strengthen health system resilience. This research contributes significantly by offering information on adopting best practices from more productive health systems, enriching knowledge about productivity in resource-constrained settings, and presenting valuable literature for future researchers. The paper's originality lies in the meticulous selection of optimal and explanatory combinations, facilitating an assessment of the technical efficiency of 46 health care systems in Africa using DEA and Tobit regression.

The aim of this study is to evaluate the technical efficiency of MNCH services across 46 African countries using a 2-stage methodology that combines DEA and Tobit regression. Specifically, this research investigates how various health system inputs and contextual explanatory variables affect the performance of MNCH services. We hypothesize that inefficiencies in MNCH services are significantly associated with health expenditures, health workforce availability, corruption, and broader socioeconomic indicators such as income inequality and out-of-pocket health costs.

The subsequent sections detail the structured literature review, methods, results, discussion, conclusions, recommendations, limitations, and future research.

Methods

Data Sources and Variables

This study included the latest data from the Global Health Observatory and WHO for 46 African countries, including information between 2005 and 2021 [5].

We selected the input, output, and explanatory variables to evaluate the accuracy of the WHO[5] statistics in describing the efficiency of MNCH. Five inputs and outputs are considered to estimate technical efficiency (Table 1).

Table . Input, output, and explanatory variables.

Variable	Description	Justification	SDG[a] link
Inputs			
Hospital beds	Hospital beds per 10,000 population	This measure indicates the capacity of the health infrastructure to provide inpatient MNCH[b] services, which is crucial for safe deliveries and emergency care.	SDG 3.c.1 (Health workforce and infrastructure)
Medical doctors	Medical doctors per 10,000 population	The availability of skilled personnel for diagnosis and treatment is crucial in reducing maternal and child mortality.	SDG 3.c.1 (Health workforce)
Nursing and midwifery personnel	Nursing and midwifery personnel per 10,000 population	Frontline providers for prenatal, delivery, and postnatal care are crucial for improving MNCH outcomes in low-resource settings.	SDG 3.c.1 (Health workforce)
Outputs			
Neonatal mortality rate	Per 1000 live births (2021)	Measures deaths within the first 28 days, indicating the effectiveness of newborn health interventions.	SDG 3.2 (Neonatal and child mortality)
Stillbirth rate	Per 1000 total births (2021)	Reflects the quality of prenatal and delivery care, highlighting gaps in maternal health services.	SDG 3.2 (Neonatal and child mortality)
Infant mortality rate	Probability of dying between birth and age 1 per 1000 live births	The infant mortality rate serves as a broad indicator of child health, reflecting factors such as vaccination, nutrition, and the effectiveness of early care.	SDG 3.2 (Neonatal and child mortality)
Births attended by skilled health personnel	Percentage	Measures access to quality maternal care, reducing risks during delivery for mothers and newborns.	SDG 3.1 (Maternal mortality)
Maternal mortality ratio	Per 100,000 live births (2020)	The maternal mortality ratio reflects the quality of maternal health services by indicating deaths from pregnancy-related causes.	SDG 3.1 (Maternal mortality)
Proportion of vaccination cards seen	Percentage	This percentage indicates the coverage and monitoring of childhood vaccinations, thereby preventing child mortality from vaccine-preventable diseases.	SDG 3.b (Access to vaccines and medicines)
Explanatory variables			
Current health expenditure	Per capita in US $ (2020)	Measures total health spending per person, reflecting investment in MNCH services like vaccinations and skilled birth attendance.	SDG 3.c (Health financing)
External health expenditure	Per capita in US $ (2021)	Captures donor funding for health, supporting MNCH programs in resource-constrained African countries.	SDG 3.c (Health financing)
Proportion of vaccination cards seen	Percentage	This indicates the coverage and monitoring of childhood vaccinations, thereby preventing child mortality from vaccine-preventable diseases.	SDG 3.b (Access to vaccines and medicines)

Variable	Description	Justification	SDG[a] link
Composite coverage index	Reproductive, maternal, newborn, and child health interventions, percentage	A composite measure of reproductive, maternal, newborn, and child health intervention coverage (eg, antenatal care, vaccinations) that summarizes health system performance.	SDG 3 (Health and well-being)

[a]SDG: Sustainable Development Goal.

[b]MNCH: maternal, newborn, and child health.

First Stage: DEA

This study used DEA to assess the technical efficiency of health care systems across 46 African countries in delivering MNCH services in the first stage.

Technical efficiency is typically measured using two methods: parametric and nonparametric [8,32-36]. A stochastic frontier production function based on a collection of explanatory variables is used in the parametric approach. The nonparametric technique, on the other hand, uses linear programming to assess the relative efficiency of DMUs by generating an ideal mix of inputs and outputs based on the best-performing unit in the collection [33,37].

Farrel introduced the DEA method [38], and Charnes et al [39] and Banker et al [40] further developed this method. The most common technique is DEA, which may be used independently or in conjunction with a secondary analysis involving the Malmquist index [41], Tobit regression [42], and correlation efficiency. Traditionally, two models are used to calculate the DEA: the CCR model developed by Charnes, Cooper, and Rhodes [39] based on the assumption of CRS and the BCC model proposed by Banker, Charnes, and Cooper based on the assumption of variable returns to scale (VRS) [40]. In the CRS model, outputs are assumed to increase proportionally with inputs, meaning that there are no economies or diseconomies of scale. This simplifies comparisons between similar-sized DMUs [39,40]. In contrast, the VRS model allows for economies and diseconomies of scale, recognizing that each DMU may have an optimal operating size. This model is better suited for comparing DMUs of different sizes [40] as it isolates pure technical efficiency from the influence of scale. DEA models can be categorized as either input-oriented or output-oriented, depending on the relationship between inputs and outputs.

DEA is a widely used method for assessing the relative efficiency of DMUs [33,37-41,43,44]. It is particularly useful when there are multiple inputs and outputs involved in the evaluation process. DEA provides a framework for DMUs and those that achieve the highest level of output given a set of inputs [12,13]. In this study, the CRS and VRS were oriented [45,46].

$$\text{Max } \sum_{r=1}^{s} u_{rk} y_{rj} + u_0 \sum_{i=1}^{m} v_{ik} x_{ik} \quad \text{constraints;}$$

$$\text{Max} \sum_{r=1}^{s} u_{rk} y_{rj} + u_0 \sum_{i=1}^{m} v_{ik} x_{ik} \leq 1 \; (j = 1, 2, 3, \ldots, n); \; v_{ik} \geq 0 \; (i = 1, 2, \ldots, s); \; (i = 1, 2, \ldots, m); \; u_0 \in R$$

The definitions and explanations of these variables are presented as follows:

- x_{ik}: input i used by DMU k
- y_{rj}: output r produced by DMU j
- v_{ki}: weight (or multiplier) assigned to input i for DMU k
- u_{rk}: weight (or multiplier) assigned to output r for DMU k
- u : a constant term (often used in affine DEA models, possibly capturing returns to scale or environmental influences)

The definitions and explanations of these indices are presented as follows:

- m: total number of inputs
- s: total number of outputs
- n: total number of DMUs being evaluated
- j: index for DMUs, where $j=1, 2, ..., n$
- i: index for inputs, where $i=1, 2, ..., m$
- r: index for outputs, where $r=1, 2, ..., s$

The following formula shows the input-oriented VRS model, with results obtained using DEAP (version 2.1) [47] in previous studies [21,40].

By comparing inefficient countries against the efficiency frontier (formed by the most efficient peers), DEA identified countries with the potential for efficiency improvement. The efficiency scores generated in this first stage were then used as a dependent variable in the Tobit regression model to analyze the influence of financial factors on inefficiency.

Second Stage: Tobit Model

To explore the financial determinants of inefficiency in MNCH service delivery, the second stage used Tobit regression, suitable for censored dependent variables—here, the DEA efficiency scores were bounded between 0 and 1.

The Tobit model was used to analyze the determinants of inefficiency in health care service delivery, specifically regarding MNCH in African countries. The Tobit model is appropriate when the dependent variable is censored, meaning some values are unobserved beyond a certain threshold [42,48]. In this case, the dependent variable is the efficiency score, bounded between 0 and 1 and censored at 0.

Data were first prepared in Microsoft Excel (Microsoft Corp) and then imported into STATA 18 (StataCorp LLC) for statistical analysis. The standard Tobit model is expressed as [42,48]:

$$y_i^* = x_i'\beta + u_i \; (i = 1, \ldots, n) \; u_i \{ y_i^*, \text{if } y_i^* > 0 \; 0, \text{if } y_i^* \leq 0 \; u_i \sim \text{IIN}(0, \sigma - 2)$$

In the formula, there is a latent random variable that is observed as y if it is positive and is otherwise observed as equal to zero and the parameter vector $\beta \in R^k$. The error I is a normal independent with a mean of zero and precision of $\sigma^2 > 0$.

We specified two Tobit regression models.

Model 1 includes the following variables: number of medical doctors (MD), number of nurses and midwives (NM), hospital bed density (HBP), current health care expenditure (CHE), and combined health care expenditure and corruption (CHEC) index.

Model 2 expands upon Model 1 by incorporating variables such as out-of-pocket costs, perceived vaccine access (PVACC), comprehensive country index (CCI), and external health contributions (EXHC).

Ethical Considerations

Our research did not require formal institutional review board or research ethics board approval as it was based entirely on secondary analysis of publicly available, anonymized data that contained no identifiable personal information and did not involve any direct interaction with human participants.

Results

Following the methodological approach outlined in the previous section, we first normalized and prepared the dataset of 46 African countries (2005 - 2021), using DEAP (version 2.1) to compute technical efficiency scores based on both CRS and VRS input-oriented DEA models. Inputs (HBP, MD, NM) and outputs (neonatal mortality rate, stillbirth rate, infant mortality rate from birth to age 1, births attended by skilled health personnel, maternal mortality per live births, PVACC) were used to estimate efficiency scores for each country. Countries achieving a score of 1 were deemed fully efficient, while scores below 1 indicated relative inefficiency. The analysis produced a frontier of best-performing countries, and inefficient countries were benchmarked against this frontier. These DEA results formed the basis for further analysis using the Tobit model to examine determinants of inefficiency.

Descriptive Statistics of the Variables Used

Mean values for the key variables to analyze descriptive statistics include 12.1 for HBP (beds per 1000 population), 3.5 for MD, and 15.2 for NM as inputs and 23.5 for neonatal mortality rate, 18.7 for stillbirth rate, 58.43 for under-5 mortality rate, and 41.5 for infant mortality rate from birth to age 1 as outputs. Explanatory variables include 75.6 for births attended by skilled health personnel, 354.2 for maternal mortality per live births, 5.7 for CHE, 134.8 for CHEC, 17.2 for EXHC, 35.3 for out-of-pocket costs, 67.7 for PVACC, and 49.1 for CCI.

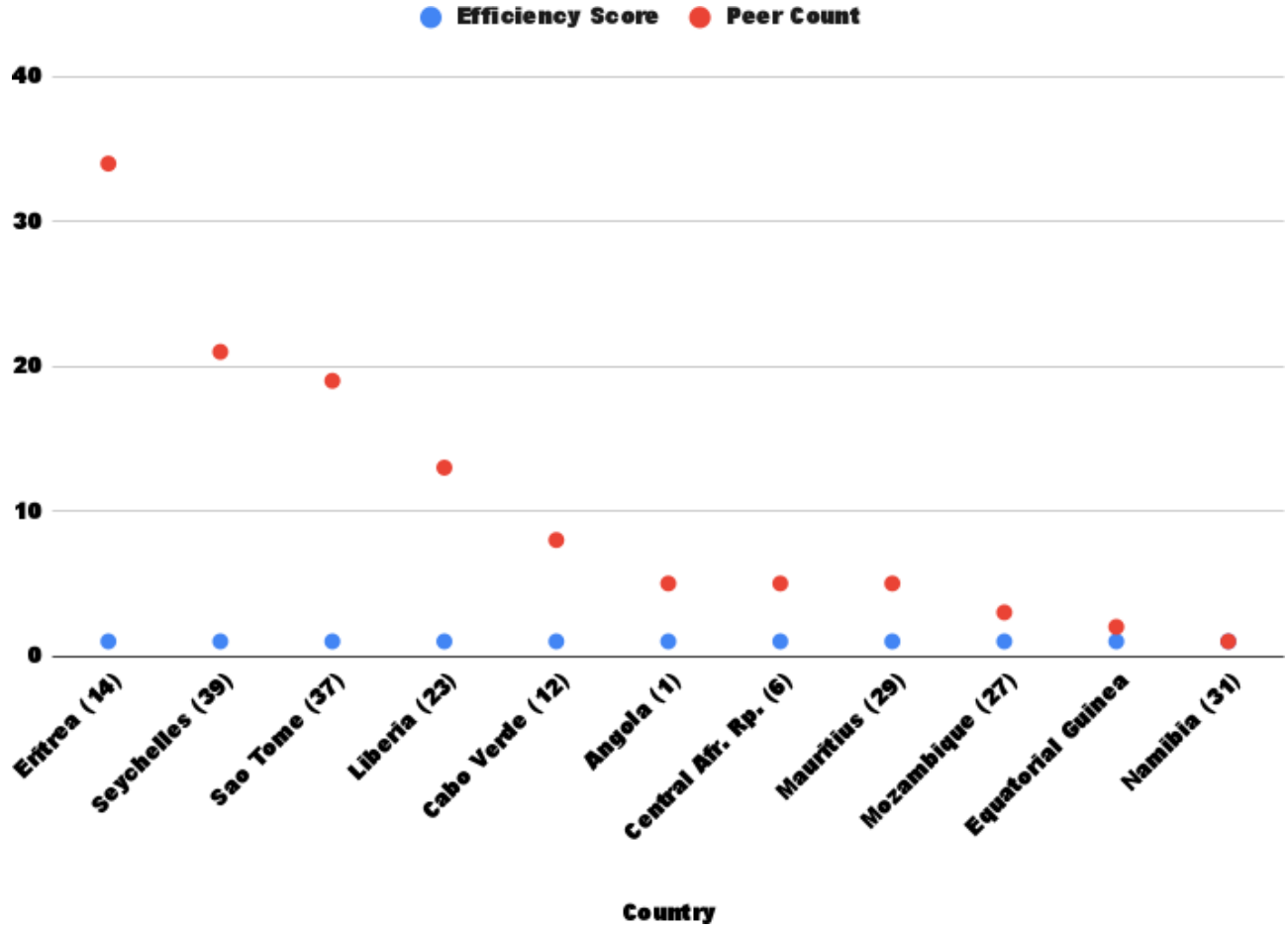
First Stage: DEA

Efficiency scores in DEA range from 0 to 1, with a score of 1 signifying that a DMU, specifically a country, is operating at the peak of efficiency, using the fewest possible inputs to achieve the observed outputs. Scores below 1 indicate relative inefficiency.

Following data preparation and input-output selection as described in the Methods, DEA was conducted using DEAP (version 2.1), applying both CRS and VRS input-oriented models. The DEA models identified which countries were efficient and how far inefficient countries were from the efficiency frontier.

Under the VRS model, 12 countries (26%) achieved full efficiency (score=1). The remaining 34 countries (74%) had efficiency scores below 1. The average technical efficiency score (technical efficiency VRS) was 0.849, indicating that, on average, countries could reduce inputs by 15.1% without compromising maternal and child health outcomes (Figure 1).

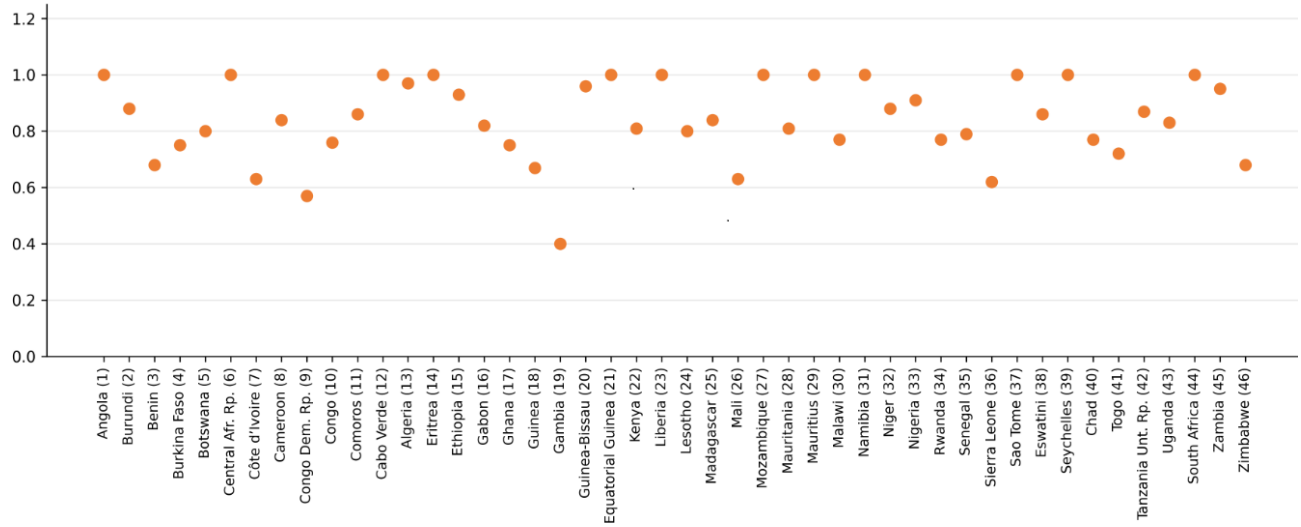
Figure 1. Input-oriented variable returns-to-scale efficiency scores. Blue represents the efficiency scores and red represents the peer count.



Gambia showed the lowest performance, with a technical efficiency VRS score of 0.403, whereas Eritrea (DMU 14) was the most frequently referenced efficient benchmark country, serving as a peer for 34 other countries. Seychelles and São Tomé also emerged as key reference points.

Additionally, the average efficiency score (technical efficiency VRS) across all countries was 0.849 for VRS, signifying that health care systems across the African continent must minimize their inputs by 15% (7 of 46 DMU) under an input orientation. Moreover, the analyzed nations displayed comparable outputs; those identified as efficient used relatively fewer resources than their inefficient counterparts. Eritrea (DMU 14) emerged as the most frequently referenced efficient country, being mentioned 34 times. From this perspective, Eritrea shares similarities with the inefficient countries in the input and output variables considered in this study. Seychelles (referenced 39 times) and São Tomé (referenced 37 times) were the next most referenced efficient countries (refer to Figure 2).

Figure 2. Peer count.



Based on World Bank income classifications, high-income countries exhibited the highest levels of technical efficiency. These nations had the highest average efficiency scores, despite making up only 13% of the sample (6 of 46 DMUs; Table 2). Countries in the middle-income category followed, with an average score of 0.86, representing 39% of the sample (18 of 46 DMUs). In contrast, countries in the lowest income classification had an average efficiency score of 0.810, comprising 43% of the total sample (16 of 46 DMUs). These results suggest that, while income level is a factor in efficiency, some lower-income countries still achieved relatively high performance due to optimal resource utilization.

Table . World Bank income classification.

World Bank income classification	Average technical efficiency	Decision-making units
Low-income (22 states)	0.810	2, 4, 6, 9, 11, 14, 15, 18, 19, 20, 23, 25, 26, 30, 32, 34, 36, 40, 41, 43, 45
Low-middle income (18 states)	0.860	1, 3, 7, 8, 10, 12, 13, 17, 22, 24, 27, 28, 33, 35, 37, 38, 42, 44, 46
High-upper-middle income (6 states)	0.940	5, 16, 21, 29, 31, 39

Second Stage: Tobit Model

Table 3 presents the results of the Tobit regressions for both models. The analysis identifies significant predictors of health care inefficiency across African countries with respect to MNCH. In Model 1, CHE was negatively associated with inefficiency scores (β=–0.0638, SE 0.0217; t=–2.94; P=.005), indicating that increased expenditure is linked to lower inefficiency. The combined effect of CHEC showed marginal significance (β=–.0040, SE 0.0024; t=–1.70; P=.096).

Table . Tobit regression. All variables tested using Tobit regression with robust standard errors.

Variable	Tobit model 1				Tobit model 2			
INF (**Inefficiency**)	Coefficient	SE	*t* test (*df*)	*P*>|*t*|	Coefficient	SE	*t* test (*df*)	*P*>|*t*|
Number of nurses and midwives	0.0009	0.0045	0.20 (40)	.84	0.0014	0.0051	0.28 (37)	.78
Number of medical doctors	–0.0068	0.0192	–0.36 (40)	.72	–0.0123	0.0168	–0.73 (37)	.47
Hospital bed density	–0.0049	0.0084	–0.58 (40)	.56	–0.0027	0.0089	–0.31 (37)	.76
Current health care expenditure	–0.0638	0.0217	–2.94 (40)	.01	–0.0811	0.0230	–3.52 (37)	.001
Combined health care expenditure and corruption	–0.0040	0.0024	–1.70 (40)	.10	–0.0019	0.0021	–0.94 (37)	.35
Perceived vaccine access	—[a]	—	—	—	–0.0000	0.0019	–0.02 (37)	.98
Comprehensive country index	—	—	—	—	0.0155	0.0041	3.75	.001
Out-of-pocket costs	—	—	—	—	—	—	—	—
External health contributions	—	—	—	—	–0.0049	0.0042	–1.16 (37)	.25
Constant	0.7136	0.1278	5.59 (40)	<.001	0.0820	0.1909	0.43 (37)	.67
Sigma	0.2684	0.0332	—	—	2.2643	0.0276	—	—
Likelihood Ratioχ^2 (*df*)	23.37 (5)	—	—	<.001	39.07 (5)	—	—	<.001
Pseudo R^2	0.5349	—	—	—	0.8943	—	—	—
Log likelihood	–10.16	—	—	—	–2.31	—	—	—

[a]Not applicable.

In Model 2, CHE remained a significant predictor (β=–.0811, SE 0.0230; t_{37}=–3.52; *P*=.001), and the CCI was positively and significantly associated with inefficiency (β=.0155, SE 0.0041; t_{37}=3.75; *P*=.001). All other variables, including MD, NM, HBP, PVACC, and EXHC, were not statistically significant (all *P* >.05).

The likelihood ratio χ^2 test was significant in both models (Model 1: χ^2_5=23.37, *P*<.001; Model 2: χ^2_5=39.07, *P*<.001), indicating that the models explained a substantial portion of the variability in inefficiency scores. Pseudo R^2 values were 0.5349 for Model 1 and 0.8943 for Model 2, suggesting better explanatory power in the second model.

Discussion

This study assessed the efficiency of MNCH systems in 46 African countries using a 2-stage approach: DEA to measure technical efficiency, followed by a Tobit regression to identify determinants of inefficiency. It found that only 26% of countries were technically efficient, while 74% were inefficient. The study also identified key determinants of inefficiency, such as CHE, corruption in health expenditure, and the CCI. Higher-income countries showed better efficiency, while low-income countries had the lowest average efficiency score. The study also found that corruption in health expenditure had a marginally significant negative association with inefficiency, while the CCI was positively and significantly associated with it. Other variables like the number of medical doctors, the number of hospitals, vaccination coverage, and out-of-pocket costs were not statistically significant.

Health systems aim to ensure equitable public access to health care services and judicious resource distribution. The responsibility for funding these requirements lies with the public. The 2030 SDGs urge governments to adopt reforms to enforce regulations in this realm, as emphasized by SDG 3. Most MNCH services rely on health care resources, and the SDGs emphasize

the need for efficient funding. This study analyzed the efficiency of MNCH services in 46 countries in Africa in the context of the SDGs, using the DEA method in the first stage and Tobit regression in the second stage.

The research findings disclose a disconcerting scenario, elucidating a substantial dissonance between the prevailing maternal and child health metrics in Africa and the specified SDGs for the year 2030.

The Atlas of African Health Statistics 2022 reveals that sub-Saharan Africa faces challenges in reducing maternal mortality and infant mortality rates. By 2030, 390 women will die in childbirth for every 100,000 live births, more than 5 times the 2030 SDG target [2]. To meet the target, Africa needs an 86% reduction from 2017 rates, which is unrealistic. The region's infant mortality rate is 72 per 1000 live births, and at the current rate of decline, 54 deaths per 1000 live births will be expected [2]. There has been some progress in key health objectives: vaccine coverage has increased, under-5 mortality has fallen by 35%, neonatal death rates dropped by 21%, and maternal mortality declined by 28% [2]. However, the region still has a long way to go, with the COVID-19 pandemic disrupting vital health services and the resurgence of vaccine-preventable disease outbreaks. Inadequate investment in health and funding for health programs are major drawbacks to meeting the SDG on health. Accelerating the agenda to meet its reduction goal will be crucial for reducing under-5 mortality to fewer than 25 deaths per 1000 live births [2].

This alarming disparity between the observed metrics and the established SDG targets underscores the considerable distance that African countries currently find themselves from realizing the objectives outlined in SDG 3. Addressing this discrepancy necessitates a comprehensive evaluation of the efficiency and various influencing variables within the MNCH domain.

The findings from the DEA analysis revealed notably low or medium efficiency for most African countries. This suggests that 22 of 46 states represent low-income countries, followed by 18 of 46 states classified as low-medium income. Eritrea was the most-referenced country. According to the Tobit model analysis, financial factors such as CHE, CCI, and CHEC harmed the inefficiency of the health system related to MNCH. These findings indicated that the health financing system suffers from profound dysfunctions, which hinders the promotion of MNCH in African countries.

According to previous studies on African countries, the performance of health systems was generally low or moderately efficient based on scores [1,20,24,25,27,29]. The WHO reported an average technical efficiency score of 0.79 across its 47 member countries in 2019 [1]. Ibrahim et al [49] assessed health care systems in sub-Saharan Africa and identified them as generally inefficient. During the analyzed period, only three provinces—in Rwanda (2014 and 2015) and in Tanzania (2015)—were found to be efficient.

The study also discovered that governance metrics, notably the rule of law and government efficacy, have a greater impact on health care system efficiency than public health spending. This implies that effective resource management is more important than the amount of money invested in health care systems in sub-Saharan African nations [49]. According to Babalola and Moodley's findings [50], less than 40% of the facilities tested were efficient. These studies reported parameters such as catchment population, facility ownership, and geography.

Arhin et al [28] discovered that by implementing best practices in instruction, management performance, expenditures on public health, external health funding, and prepayment arrangements, 30 sub-Saharan African health systems can increase UHC levels by 19% while using existing health care resources.

The overall health care efficiency in different African countries is considerable. Notably, Ghana, Sierra Leone, and Burkina Faso all recorded a low technical efficiency score in the provision of MNCH [21,24,49,51,52]. The choices of input and output variables depend on the availability of information in the reports concerning the activities of health establishments in these countries. Technical efficiency varies from one health system to another.

In Ghana's case, 78% of primary health care institutions have a low efficiency score [53]. Primary health care facilities in KwaZulu-Natal, South Africa, similarly have a 70% low technical efficiency [26]. Alhassan et al [54] found that the geographic location of the centers and their type of ownership were substantially associated with the prediction of efficiency scores rather than the quality of service. Marschall and Flessa's [51] Tobit model results in Burkina Faso demonstrated that the explanatory variables determining inefficiency in rural health care were highly related to geographical distance and other factors.

The management of African health care systems, particularly in the realm of MNCH, presents a multifaceted challenge encompassing economic, social, political, and infrastructural factors. These challenges include financial constraints, human resource shortages, infrastructure deficiencies, cultural and social barriers, governance issues, high disease burdens, inadequate health facility capacity, suboptimal utilization of health services, leakages, and corruption. Economically advanced countries such as Eritrea, Seychelles, Mauritius, Namibia, South Africa, and São Tomé exhibit efficient health systems. However, economically less developed countries encounter difficulties in providing and accessing health services due to their developmental status and less robust institutional frameworks.

A literature review revealed that countries like Ghana, Sierra Leone, and Burkina Faso all demonstrated low-efficiency scores in delivering MNCH services. It is imperative to advocate for enhanced resource allocation strategies, prioritize efficient utilization of health care resources, optimize infrastructure enhancements, invest in workforce training, and embrace technology to streamline service delivery. Health authorities are urged to consider comprehensive policy reforms aimed at addressing operational inefficiencies identified in the study. These reforms should be strategic and tailored to enhancing the overall effectiveness of health care systems in the domain of maternal and child health.

This study assessed the effectiveness of health care systems; however, its precision relied on data from the WHO, which may have overlooked key determinants influencing MNCH outcomes. Additionally, the study assumed homogeneity in production functions across diverse African countries, potentially oversimplifying variations in health care infrastructure, socioeconomic conditions, and cultural factors. Moreover, the study's focus on internal factors may have neglected external influences such as political stability and global health crises. Generalizing the findings beyond the studied nations is also risky due to the continent's heterogeneity and the dynamic nature of its efficiency. Future research could explore sustainable financing solutions for health care systems, addressing structural constraints faced by African states.

Acknowledgments

We are immensely grateful to the Ministry of Health of Morocco for their cooperation in collecting the data and making available the Annual Health Hospital Activity Reports 2021, 2022, 2023, and 2020. We are also thankful to Abdelamalek Essaadi University for facilitating the coordination of the data collection for the study.

Dr. Bouchra Merrahi, Dr. Sanaa El Achari, Dr. Nouhaila Bentatou, and Dr. Ismail El Mir contributed significantly to final data validation, methodological clarification, and thepreparation of supplementary materials required by the editorial team. They also reviewed the final version for intellectual accuracy.

Data Availability

Data are from the Ministry of Health of Morocco (2012, 2013, 2014, 2015, 2016, 2017, 2018, 2019, and 2020).

Authors' Contributions

The authors were involved in the literature review, data analysis, interpretation of the results, and drafting of the manuscript. The authors read and approved the final manuscript.

Conflicts of Interest

None declared.

References

1. Technical efficiency of health systems in the WHO African Region. World Health Organization. 2023. URL: https://www.afro.who.int/publications/technical-efficiency-health-systems-who-african-region [accessed 2025-11-07]
2. Africa's advances in maternal, infant mortality face setbacks: WHO report. World Health Organization.: Regional Office for Africa; 2025. URL: https://www.afro.who.int/news/africas-advances-maternal-infant-mortality-face-setbacks-who-report [accessed 2025-06-09]
3. Murray CJ, Frenk J. A framework for assessing the performance of health systems. Bull World Health Organ 2000;78(6):717-731. [Medline: 10916909]
4. Global health estimates: life expectancy and leading causes of death and disability. World Health Organization. 2021. URL: https://www.who.int/data/gho/data/themes/mortality-and-global-health-estimates/ [accessed 2023-12-12]
5. The Global Health Observatory Data: health and well-being. World Health Organization. 2021. URL: https://www.who.int/data/gho/data/major-themes/health-and-well-being [accessed 2023-12-18]
6. Konca M, Top M. What predicts the technical efficiency in healthcare systems of OECD countries? A two-stage DEA approach. Int J Healthc Manag 2023 Jan 2;16(1):104-119. [doi: 10.1080/20479700.2022.2077510]
7. The world health report 2000: health systems: improving performance. World Health Organization. 2000. URL: https://www.who.int/publications-detail-redirect/924156198X [accessed 2023-11-22]
8. Er-Rays Y, M'dioud M, Ait-Lemqedde H, Ezzahir M. Data envelopment analysis and Malmquist index application: efficiency of hospitals networks in Morocco. In: Ezziyyani M, Kacprzyk J, Balas VE, editors. International Conference on Advanced Intelligent Systems for Sustainable Development (AI2SD'2023): Springer Nature Switzerland; 2023, Vol. 904:13-24. [doi: 10.1007/978-3-031-52388-5_2]
9. Er-Rays Y, M'dioud M, Ait-Lemqeddem H, Ezzahiri M. Assessing efficiency maternal and child health services in Morocco: data envelopement analysis and Tobit model. Qual Quant 2024 Dec;58(6):5577-5619. [doi: 10.1007/s11135-024-01893-y]
10. Er-Rays Y, M'dioud M. Analyzing the efficiency of moroccan hospital network regions via DEA and tobit regression: assessing DEAP 2.1 software versus generative AI chatgpt 3.5. In Review. Preprint posted online on May 14, 2024. [doi: 10.21203/rs.3.rs-4369365/v1]
11. Er-Rays Y, M'dioud M. Evaluating the effectiveness of maternal, neonatal, and child healthcare in Moroccan hospitals and SDG 3: using two-stage data envelopment analysis and Tobit regression. Eval Rev 2025 Apr;49(2):343-379. [doi: 10.1177/0193841X241264863] [Medline: 39032171]

12. Chen S, Yue W, Liu N, Han X, Yang M. The progression on the measurement instruments of maternal health literacy: a scoping review. Midwifery 2022 Jun;109:103308. [doi: 10.1016/j.midw.2022.103308] [Medline: 35325678]
13. Yitbarek K, Abraham G, Adamu A, et al. Technical efficiency of neonatal health services in primary health care facilities of Southwest Ethiopia: a two-stage data envelopment analysis. Health Econ Rev 2019 Oct 27;9(1):27. [doi: 10.1186/s13561-019-0245-7] [Medline: 31656977]
14. Cetin VR, Bahce S. Measuring the efficiency of health systems of OECD countries by data envelopment analysis. Appl Econ 2016 Aug 8;48(37):3497-3507. [doi: 10.1080/00036846.2016.1139682]
15. Asandului L, Roman M, Fatulescu P. The efficiency of healthcare systems in Europe: a data envelopment analysis approach. Procedia Economics and Finance 2014;10:261-268. [doi: 10.1016/S2212-5671(14)00301-3]
16. El Husseiny IA. The efficiency of healthcare systems in the Arab countries: a two-stage data envelopment analysis approach. JHASS 2023 Aug 28;5(4):339-358. [doi: 10.1108/JHASS-10-2021-0168]
17. Hamidi S, Akinci F. Measuring efficiency of health systems of the Middle East and North Africa (MENA) region using stochastic frontier analysis. Appl Health Econ Health Policy 2016 Jun;14(3):337-347. [doi: 10.1007/s40258-016-0230-9] [Medline: 26914550]
18. Meddeb R. Efficiency of MENA Region's Health Systems: Using DEA Approach. IJISRT 2019;4(7).
19. Ahmed S, Hasan MZ, MacLennan M, et al. Measuring the efficiency of health systems in Asia: a data envelopment analysis. BMJ Open 2019 Mar 27;9(3):e022155. [doi: 10.1136/bmjopen-2018-022155] [Medline: 30918028]
20. Musoke E, Yawe BL, Ssentamu JD. The Total Factor Productivity Growth of Health Systems in African Least Developed Countries 2023. URL: https://f1000research.com/articles/12-1050 [accessed 2024-01-05] [doi: 10.12688/f1000research.135418.1]
21. Top M, Konca M, Sapaz B. Technical efficiency of healthcare systems in African countries: an application based on data envelopment analysis. Health Policy Technol 2020 Mar;9(1):62-68. [doi: 10.1016/j.hlpt.2019.11.010]
22. Novignon J, Nonvignon J. Improving primary health care facility performance in Ghana: efficiency analysis and fiscal space implications. BMC Health Serv Res 2017 Jun 12;17(1):399. [doi: 10.1186/s12913-017-2347-4] [Medline: 28606131]
23. Novignon J, Aryeetey G, Nonvignon J, et al. Efficiency of malaria service delivery in selected district-level hospitals in Ghana. Health Syst (Basingstoke) 2023;12(2):198-207. [doi: 10.1080/20476965.2021.2015251] [Medline: 37234466]
24. Ibrahim MD, Daneshvar S, Hocaoğlu MB, Oluseye OWG. An estimation of the efficiency and productivity of healthcare systems in sub-Saharan Africa: health-centred millennium development goal-based evidence. Soc Indic Res 2019 May;143(1):371-389. [doi: 10.1007/s11205-018-1969-1]
25. Kirigia JM, Asbu EZ, Greene W, Emrouznejad A. Technical efficiency, efficiency change, technical progress and productivity growth in the national health systems of continental African countries. eas 2007 Jun;23(2):19-40. [doi: 10.1353/eas.2007.0008]
26. Kirigia JM, Sambo LG, Scheel H. Technical efficiency of public clinics in Kwazulu-Natal Province of South Africa. East Afr Med J 2001 Mar;78(3 Suppl):S1-13. [doi: 10.4314/eamj.v78i3.9070] [Medline: 12002061]
27. Kirigia JM. Efficiency of Health System Units in Africa: A Data Envelopment Analysis: University of Nairobi Press; 2015. URL: https://www.researchgate.net/publication/258120669_Efficiency_of_Health_System_Units_in_Africa_A_Data_Envelopment_Analysis [accessed 2025-11-24]
28. Arhin K, Oteng-Abayie EF, Novignon J. Assessing the efficiency of health systems in achieving the universal health coverage goal: evidence from Sub-Saharan Africa. Health Econ Rev 2023 May 2;13(1):25. [doi: 10.1186/s13561-023-00433-y] [Medline: 37129773]
29. Qu J, Li A, N'Drin MGR. Measuring technology inequality across African countries using the concept of efficiency Gini coefficient. Environ Dev Sustain 2023 May;25(5):4107-4138. [doi: 10.1007/s10668-022-02236-3] [Medline: 37363029]
30. Ashmore J, Di Pietro J, Williams K, et al. A free virtual reality experience to prepare pediatric patients for magnetic resonance imaging: cross-sectional questionnaire study. JMIR Pediatr Parent 2019 Apr 18;2(1):e11684. [doi: 10.2196/11684] [Medline: 31518319]
31. Piris-Borregas S, Bellón-Vaquerizo B, Velasco-Echeburúa L, et al. Parental autonomy in the care of premature newborns and the experience of a neonatal team: observational prospective study. JMIR Pediatr Parent 2024 Aug 30;7(1):e55411. [doi: 10.2196/55411] [Medline: 39230336]
32. Er-Rays Y, M'dioud M. Evaluating the financial factors influencing maternal, newborn, and child health in Africa. arXiv. Preprint posted online on Feb 22, 2024. [doi: 10.48550/arXiv.2402.14939]
33. Asmare E, Begashaw A. Review on parametric and nonparametric methods of efficiency analysis. OABB 2018;2(2). [doi: 10.31031/OABB.2018.02.000534]
34. Er-Rays Y, Ait Lemqeddem H. Data envelopment analysis and Malmquist index application: efficiency of primary health care in Morocco and Covid-19. TURCOMAT 2021;12(5):971-983. [doi: 10.17762/turcomat.v12i5.1741]
35. Er-Rays Y, Ait Lemqeddem H. Hospital performance in Morocco and COVID-19: application of data envelopment analysis and the Malmquist index. International Journal of Accounting, Finance, Auditing, Management and Economics 2020;1(2):334-352. [doi: 10.5281/zenodo.4027715]
36. [FREE Full text]
37. Hollingsworth B. Non-parametric and parametric applications measuring efficiency in health care. Health Care Manag Sci 2003 Nov;6(4):203-218. [doi: 10.1023/a:1026255523228] [Medline: 14686627]

38. Farrell MJ. The measurement of productive efficiency. J R Stat Soc Ser A 1957;120(3):253. [doi: 10.2307/2343100]
39. Charnes A, Cooper WW, Rhodes E. Measuring the efficiency of decision making units. Eur J Oper Res 1978 Nov;2(6):429-444. [doi: 10.1016/0377-2217(78)90138-8]
40. Banker RD, Charnes A, Cooper WW. Some models for estimating technical and scale inefficiencies in data envelopment analysis. Manage Sci 1984 Sep;30(9):1078-1092. [doi: 10.1287/mnsc.30.9.1078]
41. Malmquist S. Index numbers and indifference surfaces. Trabajos de Estadistica 1953 Jun;4(2):209-242. [doi: 10.1007/BF03006863]
42. Tobin J. Estimation of relationships for limited dependent variables. Econometrica 1958 Jan;26(1):24. [doi: 10.2307/1907382]
43. Kuosmanen T, Johnson A, Saastamoinen A. Stochastic nonparametric approach to efficiency analysis: a unified framework. In: Zhu J, editor. Data Envelopment Analysis: A Handbook of Models and Methods: Springer US; 2015:191-244. [doi: 10.1007/978-1-4899-7553-9_7]
44. Bağcı H, Çil Koçyiğit S. Evaluating the decentralization of public hospitals in Turkey in terms of technical efficiency: data envelopment analysis and Malmquist index. BIJ 2023 Dec 1;30(10):4425-4460. [doi: 10.1108/BIJ-03-2021-0140]
45. Chern JY, Wan TTH. The impact of the prospective payment system on the technical efficiency of hospitals. J Med Syst 2000 Jun;24(3):159-172. [doi: 10.1023/a:1005542324990] [Medline: 10984870]
46. Sherman HD, Zhu J. Service Productivity Management: Improving Service Performance Using Data Envelopment Analysis (DEA): Springer Science & Business Media; 2006. [doi: 10.1007/0-387-33231-6]
47. Coelli T. A guide to DEAP version 21: a data envelopment (computer) program. URL: https://www.owlnet.rice.edu/~econ380/DEAP.PDF [accessed 2025-11-07]
48. Amemiya T. Tobit models: a survey. J Econom 1984 Jan;24(1-2):3-61. [doi: 10.1016/0304-4076(84)90074-5]
49. Ibrahim MD. Efficiency and productivity analysis of maternal and infant healthcare services in Sub-Saharan Africa. Int J Health Plann Manage 2023 Nov;38(6):1816-1832. [doi: 10.1002/hpm.3705] [Medline: 37674352]
50. Babalola TK, Moodley I. Assessing the efficiency of health-care facilities in sub-Saharan Africa: a systematic review. Health Serv Res Manag Epidemiol 2020;7:2333392820919604. [doi: 10.1177/2333392820919604] [Medline: 32426420]
51. Marschall P, Flessa S. Efficiency of primary care in rural Burkina Faso. A two-stage DEA analysis. Health Econ Rev 2011 Jul 20;1(1):5. [doi: 10.1186/2191-1991-1-5] [Medline: 22828358]
52. Marschall P, Flessa S. Assessing the efficiency of rural health centres in Burkina Faso: an application of Data Envelopment Analysis. J Public Health 2009 Apr;17(2):87-95. [doi: 10.1007/s10389-008-0225-6]
53. Akazili J, Adjuik M, Jehu-Appiah C, Zere E. Using data envelopment analysis to measure the extent of technical efficiency of public health centres in Ghana. BMC Int Health Hum Rights 2008 Nov 20;8(1):11. [doi: 10.1186/1472-698X-8-11] [Medline: 19021906]
54. Alhassan RK, Nketiah-Amponsah E, Akazili J, Spieker N, Arhinful DK, Rinke de Wit TF. Efficiency of private and public primary health facilities accredited by the National Health Insurance Authority in Ghana. Cost Eff Resour Alloc 2015;13:23. [doi: 10.1186/s12962-015-0050-z] [Medline: 26709349]

Edited by F Wu; submitted 19.04.24; peer-reviewed by M Mahundi, T Olorunyomi; revised version received 09.06.25; accepted 19.06.25; published 28.11.25.

Please cite as:
Er-Rays Y, M'dioud M, Ait-Lemqeddem H, El Moutaqi B
Evaluating the Financial Factors Influencing Maternal, Newborn, and Child Health in Africa: Tobit Regression and Data Envelopment Analysis
JMIRx Med 2025;6:e59703
URL: https://xmed.jmir.org/2025/1/e59703
doi:10.2196/59703

Improving Tuberculosis Detection in Chest X-Ray Images Through Transfer Learning and Deep Learning: Comparative Study of Convolutional Neural Network Architectures

Alex Mirugwe[1], BSc, MSci; Lillian Tamale[2], PhD; Juwa Nyirenda[3], PhD

[1]School of Public Health, Makerere University, Kawalya Kaggwa Close, Plot 20A, Kampala, Uganda

[2]Faculty of Science and Technology, Victoria University, Kampala, Uganda

[3]Department of Statistical Sciences, University of Cape Town, Cape Town, South Africa

Corresponding Author:
Alex Mirugwe, BSc, MSci
School of Public Health, Makerere University, Kawalya Kaggwa Close, Plot 20A, Kampala, Uganda

Related Articles:

Companion article: https://www.medrxiv.org/content/10.1101/2024.08.02.24311396v1

Companion article: https://med.jmirx.org/2025/1/e 77171

Companion article: https://med.jmirx.org/2025/1/e 77174

Companion article: https://med.jmirx.org/2025/1/e 77221

Abstract

Background: Tuberculosis (TB) remains a significant global health challenge, as current diagnostic methods are often resource-intensive, time-consuming, and inaccessible in many high-burden communities, necessitating more efficient and accurate diagnostic methods to improve early detection and treatment outcomes.

Objective: This study aimed to evaluate the performance of 6 convolutional neural network architectures—Visual Geometry Group-16 (VGG16), VGG19, Residual Network-50 (ResNet50), ResNet101, ResNet152, and Inception-ResNet-V2—in classifying chest x-ray (CXR) images as either normal or TB-positive. The impact of data augmentation on model performance, training times, and parameter counts was also assessed.

Methods: The dataset of 4200 CXR images, comprising 700 labeled as TB-positive and 3500 as normal cases, was used to train and test the models. Evaluation metrics included accuracy, precision, recall, F_1-score, and area under the receiver operating characteristic curve. The computational efficiency of each model was analyzed by comparing training times and parameter counts.

Results: VGG16 outperformed the other architectures, achieving an accuracy of 99.4%, precision of 97.9%, recall of 98.6%, F_1-score of 98.3%, and area under the receiver operating characteristic curve of 98.25%. This superior performance is significant because it demonstrates that a simpler model can deliver exceptional diagnostic accuracy while requiring fewer computational resources. Surprisingly, data augmentation did not improve performance, suggesting that the original dataset's diversity was sufficient. Models with large numbers of parameters, such as ResNet152 and Inception-ResNet-V2, required longer training times without yielding proportionally better performance.

Conclusions: Simpler models like VGG16 offer a favorable balance between diagnostic accuracy and computational efficiency for TB detection in CXR images. These findings highlight the need to tailor model selection to task-specific requirements, providing valuable insights for future research and clinical implementations in medical image classification.

(JMIRx Med 2025;6:e66029) doi:10.2196/66029

KEYWORDS

tuberculosis detection; tuberculosis; TB; chest x-ray classification; diagnostic imaging; radiology; medical imaging; convolutional neural networks; data augmentation; deep learning; early warning; early detection; comparative study

Introduction

Background

Tuberculosis (TB) remains one of the leading infectious diseases worldwide, affecting an estimated one-third to one-fourth of the global population with the bacillus *Mycobacterium tuberculosis*, the causative agent of TB [1]. In 2019, it was estimated that over 10 million individuals globally contracted TB; yet, only 71% were detected, diagnosed, and reported through various countries' national TB programs, leaving approximately 29% of cases unreported [2]. According to the World Health Organization's (WHO's) 2023 TB report, TB was identified as the second most common cause of death among infectious diseases [3]. Furthermore, the global incidence rate of TB remains alarmingly high at approximately 133 new cases per 100,000 people annually. This situation underscores the need for prompt, effective, and affordable screening and treatment strategies to meet the WHO's ambitious goals of reducing TB incidence by 80%, decreasing TB mortality by 90%, and eliminating catastrophic financial burdens on families affected by TB by 2030 [4].

The WHO advised member countries to proactively conduct TB screening and detection, especially within the high-risk groups, taking into account their unique epidemic scenarios and financial levels [5]. While bacteriological tests, including sputum cultures, sputum smears, and molecular diagnostics, are considered the gold standard for identifying active TB cases, their applicability on a large scale, particularly among high-risk populations, is not feasible [6]. This limitation is due to the methods being resource-intensive, logistically challenging, and associated with prolonged turnaround times [7]. As a result, chest radiography has become the most prevalent method for early TB detection [8]. However, in countries with limited resources, which also bear the highest TB burden, the availability of chest radiography screenings remains inadequate, primarily due to a shortage of radiologists [6].

In recent years, significant advancements have been made in leveraging artificial intelligence (AI), particularly through machine learning and deep learning techniques, for analyzing chest x-ray (CXR) images to differentiate between TB-positive and TB-negative images [9-15]. This innovation has enabled individuals without radiology expertise to conduct TB screening tests, presenting a significant shift in diagnostic approaches. These technologies have shown promising results, to the extent of outperforming radiologists in the interpretation of CXR images [14,15]. Despite this progress, the adoption of AI-based TB detection in low-income countries faces limitations, including a lack of computational resources, inconsistent data quality, and the need for models tailored to diverse clinical and demographic contexts. Addressing these challenges is critical to ensuring the scalability and utility of AI-driven diagnostic tools in these settings.

This research investigates the effectiveness of different convolutional neural network (CNN) architectures in classifying TB in CXR images. We compare and evaluate the performance of popular CNN models, including Residual Network (ResNet), Inception, and Visual Geometry Group (VGG), and examine the impact of different hyperparameters on classification accuracy. The choice of these architectures is motivated by gaps in existing literature, where limited studies compare the performance of advanced CNN models on larger, diverse datasets. Additionally, we explore the impact of transfer learning and data augmentation techniques, providing insights into their role in optimizing model performance.

To the best of our knowledge, this study is the first to use a larger and more diverse dataset and conduct a comprehensive comparison of the latest CNN architectures, including ResNet101, ResNet152, and Inception-V2, assessed across different parameters. The research aims to address the following questions: (1) How does the choice of CNN architecture affect the classification performance? (2) What is the optimal hyperparameter configuration for each CNN architecture? (3) Can transfer learning be leveraged to improve classification accuracy? (4) How does incorporating data augmentation techniques impact the model's performance compared to training solely on real images?

The rest of the paper is organized as follows. In the Related Work section, we present the literature review, which provides an overview of the current state of research in the field. This is followed by the Methods section, where we describe the deep learning models used in this research along with the techniques for improving training time, such as transfer learning. We also describe the data and analysis procedures used in our study, such as data augmentation to mitigate against imbalance. Next, we present the results of our analysis, including any findings. Finally, we discuss the implications of our results, conclude with a summary of our main findings, and suggest areas for future research.

Related Work

Research in the field of medical imaging, particularly in automating the screening and identification of TB from CXR images, has progressed significantly. Initial investigations explored traditional machine learning techniques, including support vector machines [16,17], decision trees [18,19], random forests [20,21], and extreme gradient boosting [22,23], among others. However, recent advancements have shifted focus toward deep learning methods, such as CNNs, which have demonstrated promising results in image classification comparable to those of radiologists [13-15,24]. Below, we review some of the recent studies that have used deep learning approaches for detecting TB in CXR images.

Hooda et al [13] proposed a 19-layer CNN architecture for detecting TB, consisting of 7 convolutional layers, 7 rectified linear unit (ReLU) layers, 3 fully connected layers, and 2 dropouts layers. The model was trained on a dataset of 800 CXR images, each resized to 224×224 pixels. Using the Adam optimizer, the study achieved notable results, with an overall accuracy of 94.73% and a validation accuracy of 82.09%. Although these results are impressive, the authors identified potential areas for further improvements. They suggested investigating the impacts of data augmentation and transfer learning on the model's performance, highlighting avenues for future research enhancements and potential increases in accuracy.

Ojasvi et al [25] developed a classification algorithm for CXR images of potential patients with TB, aiming to improve upon existing models [26]. To mitigate against dataset imbalances and improve model reliability, they combined the NIH Chest X-ray Dataset, China-Shenzhen Chest X-ray Database, and Montgomery County Chest X-ray Database to train and fine-tune their model. By implementing coarse-to-fine transfer learning and extensive data augmentation techniques, they achieved a remarkable accuracy of 94.89% compared to the accuracy of 89.6% achieved by Cao et al [26]. However, the study acknowledges the challenge of maintaining equivalent precision across CXR images obtained in varied settings, as the model was specifically trained for the Chinese dataset.

Panicker et al [27] introduced a novel 2-stage detection method for TB bacilli, using image binarization and CNN classification to analyze microscopic sputum smear images. The method was evaluated on a diverse dataset of 22 images, and the model demonstrated high effectiveness, achieving a recall rate of 97.13%, a precision of 78.4%, and an F_1-score of 86.76%. However, the study noted that the model's ability to accurately detect overlapping bacilli was limited. In the same year, Stirenko et al [28] explored the application of lung segmentation in CXR images and data augmentation to enhance TB detection from CXR images. Their study highlights the critical role of preprocessing, including lung segmentation and data augmentation, in addressing overfitting issues and improving the effectiveness of computer-aided diagnosis systems in TB identification, particularly when working with limited datasets.

The study by Kazemzadeh et al [15] developed a deep learning algorithm for detecting active pulmonary TB from CXR images. The algorithm was trained and validated on a dataset comprising 165,754 images from 22,284 patients from 10 different countries. The algorithm's performance was compared to that of 14 radiologists on datasets from 4 countries, including a cohort from a South African mining population. It achieved an area under the receiver operating characteristic curve (AUC-ROC) of 0.89, with superior sensitivity (88% vs 75%; P=.05) and comparable specificity (79% vs 84%) to radiologists, demonstrating its potential for TB screening in resource-limited settings. Another study by Nijiati et al [29] used a 3D ResNet-50 CNN architecture to differentiate active from nonactive pulmonary TB using computed tomography images. This study, similar to that of Kazemzadeh et al [15], reported high diagnostic accuracy and efficiency, outperforming conventional radiological methods in terms of speed and precision.

In their 2019 study, Meraj et al [30] used CNN architectures such as VGG16, VGG19, ResNet50, and GoogLeNet to automate the detection of TB manifestations in CXRs using 2 public TB image datasets [31]. Their findings showed that the VGG16 model outperformed other architectures in terms of accuracy and AUC-ROC. However, the study was limited by its reliance on small and unbalanced datasets, raising questions about the generalizability of the results. In contrast, our research builds upon and extends the work of Meraj et al [30] by incorporating a larger and more diverse dataset. We also explore the diagnostic capabilities of more advanced CNN architectures, including ResNet101, ResNet152, and Inception-V2, to assess their effectiveness in TB detection. This approach aims to provide a more comprehensive understanding of how recent deep learning advancements can be leveraged for more accurate TB diagnosis in varied clinical settings. The Methods section details the methodological framework to achieve these objectives.

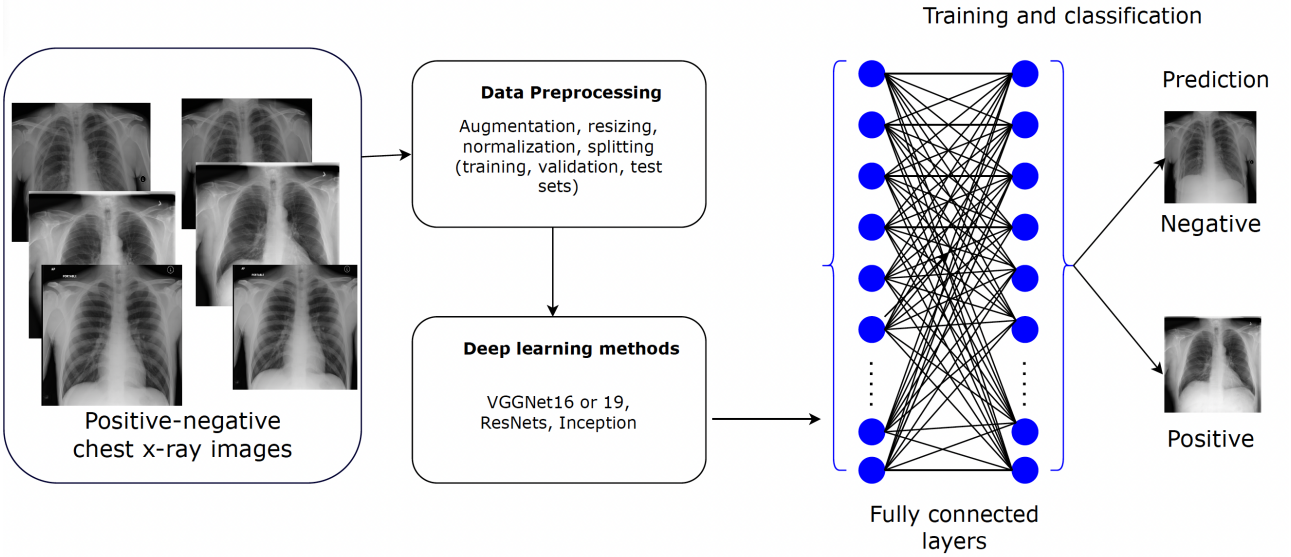
Methods

In this section, we provide a comprehensive overview of the methodologies used in our study, including the dataset and preprocessing, data normalization, data augmentation, the application of transfer learning methods, the architecture of CNNs used, and the evaluation metrics adopted to assess the performance of the models.

Implementation Overview

The implementation framework illustrated in Figure 1 starts with the acquisition of a well-defined dataset, followed by comprehensive data preprocessing, which includes data augmentation, resizing, normalization, and partitioning into training, validation, and test sets. Subsequently, we embark on the development of various deep learning models. These models undergo extensive training and evaluation against different hyperparameters and evaluation metrics to accurately predict and classify CXR images into positive or negative cases of TB.

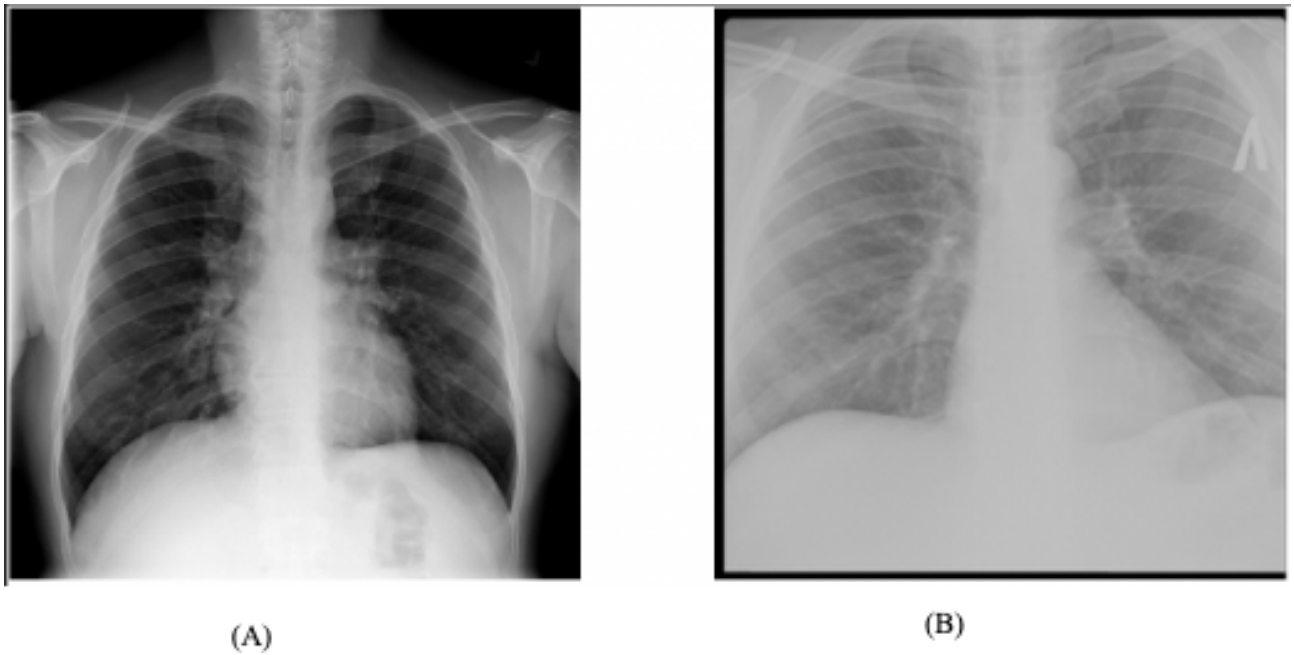
Figure 1. The implementation flow of the deep learning classification methodology. ResNet: Residual Network; VGG: Visual Geometry Group.



Dataset

The dataset used in this research comprises 4200 CXR images sourced from a public Kaggle data repository. The dataset was compiled through a collaborative effort between researchers from Qatar University (Doha, Qatar) and the University of Dhaka (Bangladesh) and collaborators from Malaysia. They worked closely with medical professionals from the Hamad Medical Corporation (Doha, Qatar) and various health care institutions in Bangladesh. The dataset consists of 700 CXR images indicative of TB and 3500 CXR images classified as normal, with all images having a resolution of 512×512 pixels [32]. This composition provides a substantial foundation for evaluating the effectiveness of CNN models in the detection of TB from CXR images. Figure 2 presents some of the images from the dataset.

Figure 2. The chest x-ray sample images. (**A**) Tuberculosis-negative and (**B**) tuberculosis-positive.



Preprocessing

To optimize the performance and efficiency of our models, we implemented key preprocessing techniques, specifically data normalization and augmentation, before training the models.

Data Normalization

In the preprocessing stage of image analysis, normalization is a critical step to standardize the input data, facilitating the model's learning process. This study applies normalization to CXR images, which initially possess pixel intensity values in the range of 0 to 255, common for grayscale images [33]. The goal of normalization is to adjust these intensity values to a standardized scale that improves computational efficiency and model convergence during training. The normalization process is mathematically represented as follows:

$$I' = \frac{I - I_{min}}{I_{max} - I_{min}} \quad (1)$$

where I represents the original pixel intensity of the image, I_{min} and I_{max} are the minimum and maximum possible intensity values in the original image, respectively, and I is the normalized pixel intensity.

For grayscale images, I_{min}=0 and I_{max}=255. This equation effectively rescales the pixel intensity values to the range (0-1), making the input data more suitable for processing by the neural network layers. This normalization technique is advantageous because it ensures that each input parameter (pixel, in this case) contributes equally to the analysis, preventing features with initially larger ranges from dominating the learning process [34]. It also helps to stabilize the gradient descent optimization algorithm by maintaining a consistent scale for all gradients [35]. Previous studies have shown that normalization significantly improves convergence rates and ensures model stability, particularly in image classification tasks involving deep learning [34,35].

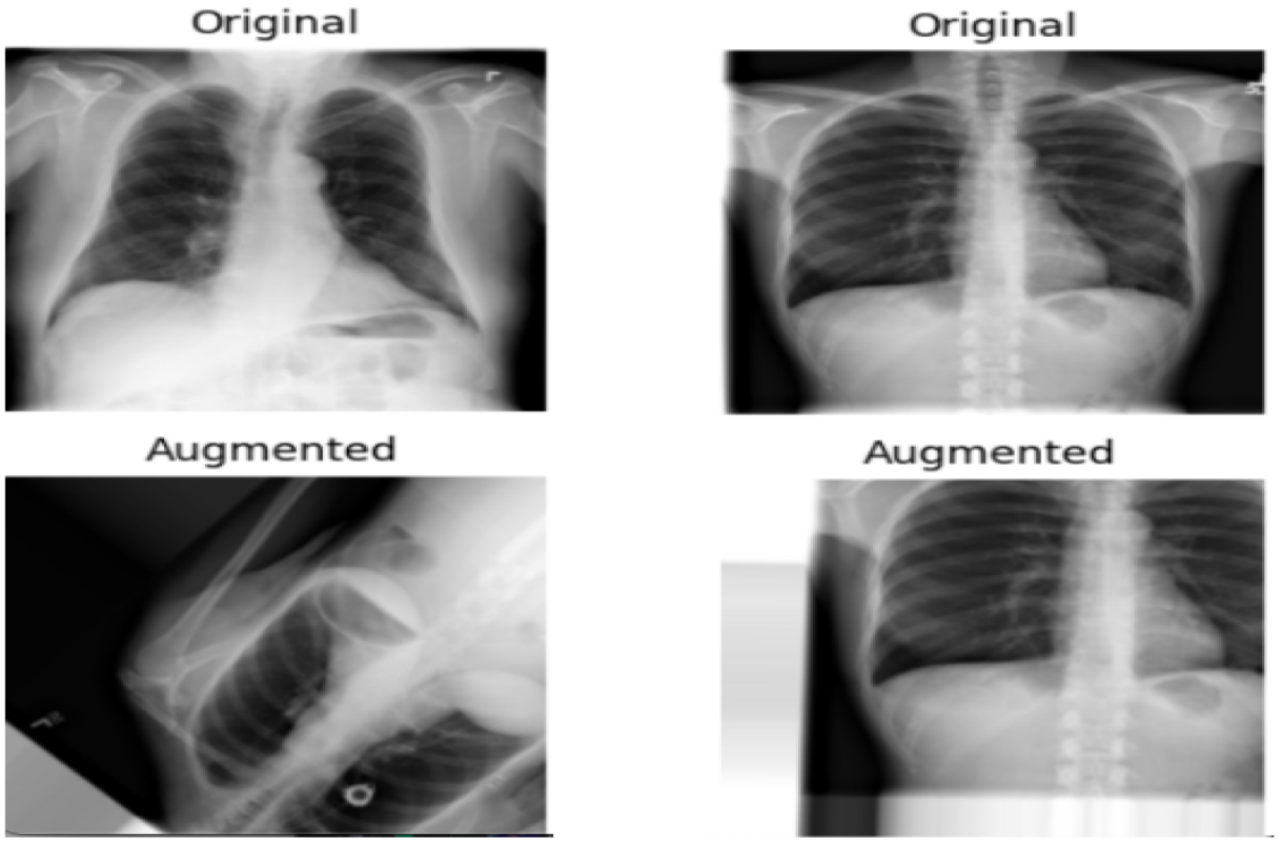
Data Augmentation

Data augmentation represents a powerful regularization strategy designed to artificially increase the dataset through label-preserving transformations, thereby incorporating more invariant examples into the training set [36]. This approach, characterized by its computational efficiency, has been previously used to reduce overfitting when training CNNs, such as in the ImageNet Large-Scale Visual Recognition Challenge (ILSVRC), where it contributed to achieving state-of-the-art results [37]. This method enhances the robustness and generalizability of deep learning models by exposing them to a wider array of variations, simulating real-world variability.

In our study, to address the imbalance between TB-positive and TB-negative images and to introduce different variations, we randomly augmented 210 (30%) TB-positive images and 175 (5%) TB-negative images. The data augmentation techniques applied included random rotation within a range of 0 to 60 degrees, random width and height shifts of up to 0.2 times the image size, and random zooming of up to 0.2 times the original size, alongside horizontal and vertical flipping. To manage the newly created pixels from such transformations, a “fill mode” strategy was used, ensuring integrity and consistency in the augmented images. These augmentations were performed using Keras’s ImageDataGenerator, a comprehensive data augmentation suite [38].

While data augmentation techniques are widely adopted in deep learning research, our implementation aligns with prior studies that highlight their utility in addressing dataset imbalance and improving model generalization in medical imaging tasks [36,37]. Additionally, the augmentation strategy in this study was tailored to reflect the variability commonly observed in real-world CXR data, enhancing the robustness of our models. Figure 3 shows a sample of real images and their corresponding augmented outputs.

Figure 3. Sample of real and corresponding augmented images.



Transfer Learning

Transfer learning is a machine learning technique where a model developed for a specific task is repurposed as the starting point for a model on a second, related task [39]. This technique leverages the knowledge gained during the initial training phase in one domain to enhance learning in another potentially unrelated domain. It operates under the principle that information learned in one context can be exploited to accelerate or improve the optimization process in another, essentially allowing for the transfer of learned features and patterns across different but related problems [39].

In this study, we propose an implementation that capitalizes on the transfer learning paradigm by using pretrained models such as Inception-V3, ResNet (50, 101, and 152), and VGG (16 and 19), which were initially trained on the ImageNet dataset [37]. This adaptation involves fine-tuning and customizing the models' last layers to suit our classification task, effectively tailoring the robust, prelearned representations of the ImageNet dataset to recognize and interpret the specific patterns and anomalies associated with TB in CXR images.

We opted for transfer learning over training models from scratch due to its significant advantages, particularly in the context of medical imaging. Training deep learning models from scratch requires large datasets, extensive computational resources, and longer training times. These requirements often pose challenges in health care–related research, especially when working with relatively small or domain-specific datasets like CXRs. Transfer learning allows us to leverage the rich feature representations of pretrained models while reducing training time and computational demands. Furthermore, studies have shown that transfer learning enhances model performance in medical imaging tasks by effectively repurposing features learned from general image datasets like ImageNet to domain-specific tasks [37,39].

CNN Architectures

In the next subsections, we provide a brief description of the VGG and ResNet families of CNN architectures as well as the Inception ResNet architecture that is considered in this study.

VGGNet

Introduced by Simonyan and Zisserman from the University of Oxford's Visual Geometry Group in 2014, the VGGNet architecture marked a significant milestone in the field of deep learning [40]. Known for its outstanding performance in the ILSVRC of that year, VGGNet is characterized by its use of 3×3 filters in all convolutional layers, simulating the effects of larger receptive fields. This architecture is available in 2 variants, VGG16 and VGG19, differing in depth and the number of layers, with VGG19 being the deeper model.

In our research, we used both the VGG16 and VGG19 architectures to train models on datasets consisting of solely real CXR images and a combination of augmented and real images. This approach aimed to assess the impact of incorporating augmented images on the performance of these 2 architectures. Images were resized to 256×256 pixels before being input into the networks. We extended the architectures by adding a flattening layer, followed by a dense layer of 512 neurons with a ReLU activation function and a dropout layer with a dropout rate of 0.2 to mitigate overfitting. A softmax activation function was used in the output layer for binary classification. We used the Adam optimizer with the binary cross-entropy loss function for optimization. The training was conducted over 15 epochs with a batch size of 32 for both models. This rigorous approach ensured that both architectures could classify between TB-positive and TB-negative CXR images accurately.

ResNet

He et al [41] introduced the deep residual network (ResNet) architecture in their 2016 seminal paper. This architecture greatly improved the performance of deep neural networks and went on to win the Common Objects in Context object detection challenge and the 2015 ILSVRC. To date, several variants of the ResNet architecture exist, including ResNet50, ResNet101, and ResNet152, which vary in depth and number of layers. ResNet architectures are very deep models [41,42]. The core idea behind ResNet is the use of residual connections, also known as shortcuts, which bypass 1 or more layers. By resolving the vanishing gradient issue, these shortcuts maintain the gradient flow across the network and facilitate the training of much deeper networks [41].

The CXR images in this study were classified using the ResNet50, ResNet101, and ResNet152 architectures. We added 3 more layers to the ResNet50 model, 2, each with 256 units and 1 with 512 units, using batch normalization and ReLU activation in each layer. To reduce overfitting, dropout layers were added with dropout rates of 0.3, 0.25, and 0.2, respectively. The binary cross-entropy loss function was used to compile the model, while the Adam optimizer was used to optimize the model at a learning rate of 0.001. Two units with a softmax activation function made up the output layer, which classified the images as either TB-positive or TB-negative. Training for this model involved 16 batch sizes and 100 epochs.

ResNet101 was trained using the same settings as ResNet50, as preliminary training showed that the same parameter values used for ResNet50 also yielded optimal results for the ResNet101 architecture. For ResNet152, a selective fine-tuning approach was adopted, where only the last 10 layers of the network were trainable, enhancing the model's focus on more feature-specific adjustments in the later stages of the network. This model shared the augmentation layers of ResNet50 but was trained for only 50 epochs, incorporating a learning rate scheduler, ReduceLROnPlateau, which adjusted the rate based on the validation loss with a factor of 0.1, patience of 5, and a minimum learning rate of 1×10^{-6}, thereby optimizing the training dynamics. The details of the models' configuration are shown in Table 1.

Table . Training hyperparameters of ResNet[a] models.

Hyperparameter	ResNet50	ResNet101	ResNet152
Layers, n	53 (50 base +3 extra)	104 (101 base +3 extra)	155 (152 base +3 extra)
Units per layer	256, 256, 512	256, 256, 512	256, 256, 512
Activation	ReLU[b]	ReLU	ReLU
Batch normalization	Yes	Yes	Yes
Dropout rate	0.3, 0.25, 0.2	0.3, 0.25, 0.2	0.3, 0.25, 0.2
Optimizer	Adam	Adam	Adam
Learning rate	0.001	0.001	Variable (ReduceLROnPlateau)
Loss function	Binary cross-entropy	Binary cross-entropy	Binary cross-entropy
Training epochs	100	100	50
Batch size	16	16	16

[a]ResNet: Residual Network.

[b]ReLU: rectified linear unit.

Inception-ResNet

The Inception networks, introduced by Szegedy et al [43], have greatly advanced the field of CNN, as they have achieved state-of-the-art performance in a number of computer vision problems [43-45]. The original Inception-V1, also known as GoogLeNet, was first introduced in 2014 and won the ILSVRC of that year. The architecture introduced a novel approach of using multiple convolutional filter sizes in parallel, allowing the network to capture various spatial features of different scales with improved use of computing resources [43].

In this study, we used Inception-ResNet-V2 architecture, a hybrid model that combines the benefits of both the Inception and residual networks. This hybrid approach enables the architecture to learn more complex features with improved training stability and faster convergence [43]. The Inception-ResNet-V2 also leverages residual connections to skip certain layers during training, which helps it improve gradient flow, accelerate training times, and reduce the likelihood of vanishing gradient problems in deep networks [46]. We selected Inception-ResNet-V2 due to its demonstrated state-of-the-art results in several medical imaging tasks [45].

For our implementation, the Inception-ResNet-V2 architecture was initialized with weights pretrained on the ImageNet dataset. Similar to our approach with the ResNet152 model, all layers except the last 10 were frozen to retain the pretrained features from ImageNet. The last 10 layers were set to be trainable, enabling the model to learn specific features from the CXR images. We added 3 new layers: 2 with 256 units each and 1 with 512 units, all using ReLU activations and batch normalization. Each of these layers was followed by dropout layers with rates of 0.4, 0.35, and 0.3, respectively, to introduce nonlinearity and reduce overfitting. The final output layer consisted of 2 units with a softmax activation function for binary classification. The model was then compiled using binary cross-entropy as the loss function and the Adam optimizer with a learning rate of 0.0001. Training was conducted for 50 epochs with a batch size of 16.

The parameters used in the training of all these CNN architectures, including dropout rates, learning rates, batch sizes, and the number of epochs, were determined through a rigorous iterative process of experimentation. This approach involved fine-tuning each parameter to optimize model performance while avoiding overfitting. The configurations presented reflect the parameter values that consistently yielded good performance across the different architectures.

Evaluation Metrics

The performance of the CNN architectures in classifying CXR images into TB-positive and TB-negative categories was assessed using several standard performance metrics, including accuracy, precision, recall, F_1-score, and the AUC-ROC. Each metric provides unique insights into the model's classification abilities, considering both the true and false predictions.

Accuracy

This metric measures the proportion of true positive (TP) and true negative (TN) results among the total number of cases examined:

(2)Accuracy=TP+TNTP+TN+FP+FN

where TP is the number of TB-positive images that are correctly identified as TB-positive by the model, TN is the number of TB-negative images that are correctly identified as TB-negative by the model, FP (false positives) is the number of TB-negative images that are incorrectly identified as TB-positive by the model, and FN (false negatives) is the number of TB-positive images that are incorrectly identified as TB-negative by the model.

Precision

Also known as positive predictive value, precision is the ratio of correctly identified TB cases to all cases that were diagnosed as TB by the model. It measures the model's accuracy in diagnosing a patient with TB when the model predicts the disease. High precision indicates a low rate of false TB diagnoses. Mathematically, it is defined as:

(3)Precision=TPTP+FP

Recall

Recall, or sensitivity, is especially critical in medical diagnostics, as it quantifies the model's ability to correctly identify all actual TB cases. It represents the proportion of actual TB cases that were correctly identified by the model and aims to minimize the risk of missing a true TB case. It is computed as:

(4)Recall= TPTP+FN

F_1-Score

The F_1-score is the harmonic mean of precision and recall, providing a single measure that balances both the FP and FN. In TB diagnosis, it is particularly useful because it creates a balance between precision (minimizing false TB diagnoses) and recall (minimizing missed TB diagnoses), which is crucial for medical screening tests. It is defined as:

(5)F1=2 ×Precision ×RecallPrecision+Recall

AUC-ROC

The AUC-ROC measures a model's ability to discern between positive and negative classes. In the context of our problem, that specifically refers to distinguishing between TB-positive and TB-negative CXR images. The AUC-ROC is a plot of the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings. The AUC-ROC provides an aggregated measure of the model's performance across all classification thresholds, with a value of 1 representing a perfect model and a value of 0.5 representing a model with no discriminatory power. The approximate AUC-ROC is calculated by using the following formula:

(6)AUC ≈∑i=1n(FPRi−FPRi−1)×(TPRi+ TPRI−1)2

where i is the current data point or threshold, FPR_i and TPR_i are the false positive and true positive rates at the ith threshold, respectively, and n is the number of data points or thresholds used to calculate the AUC-ROC. Each term in the sum represents the area of a trapezoid, where $(FPR_i - FPR_{i-1})$ is the base of the trapezoid and $(TPR_i + TPR_{i-1})/2$ is the average height of the trapezoid. The formula calculates the AUC-ROC by summing the areas of trapezoids formed by connecting consecutive points on the AUC-ROC.

Computational Environment

The implementation and findings of this study were based on using the Keras 3.3.3 and TensorFlow 2.16.1 frameworks. The experiments were conducted on a single GPU MSI GL75 Leopard 10SFR laptop with 32 GB of RAM and an 8 GB NVIDIA GEFORCE RTX 2070 GDDR6 card. The system was operated using the CUDA 12.1 and cuDNN SDK 8.7.0 platforms to ensure efficient GPU acceleration and deep learning model training.

These methodological choices, including dataset selection, preprocessing techniques, CNN architectures, and model evaluation techniques, were designed to ensure a rigorous and comprehensive analysis of CNN performance for TB detection. The results of these analyses are presented in the following section.

Ethical Considerations

This study used a publicly available, deidentified dataset from Kaggle. As such, it did not require institutional review board approval. The dataset does not contain any personally identifiable information, and informed consent was not applicable. No participants were directly involved in this study, and no compensation was provided.

Results

Overview

The study aimed to analyze and compare the performance of various CNN architectures, including VGG16, VGG19, ResNet50, ResNet101, ResNet152, and Inception-ResNet-V2, in classifying CXR images as either TB-positive or TB-negative. Additionally, we also investigated whether data augmentation could further improve the classification performance of these models by comparing the performance of models trained on only real images versus those trained on a combination of real and augmented data. We went further to examine the training time and the number of parameters for each architecture to understand the computational efficiency and resource demands for each model. This analysis is important for practical implementation, particularly in resource-constrained settings where training time and computational costs are significant considerations. By evaluating these parameters, we aimed to identify models that not only perform well but also offer a balanced trade-off between accuracy and efficiency, making them suitable for real-world applications in diverse health care environments.

Table 2 summarizes the performance of CNN architectures across accuracy, precision, recall, and F_1-score, highlighting the impact of training on real images versus a combination of real and augmented data. Table 3 shows the performance of these models when evaluated using the AUC-ROC score metric. It was observed that the VGG16 outperformed all other architectures across all metrics, with an accuracy of 99.4%, precision of 97.9%, recall of 98.6%, F_1-score of 98.3%, and area under the curve of 98.25%. Its performance was superior consistently, irrespective of whether the models were trained with or without data augmentation.

Table . Evaluation of convolutional neural network (CNN) architectures across key evaluation metrics[a].

Architecture	Accuracy (%)	Precision (%)	Recall (%)	F_1-score (%)
VGG16[b]	99.4	97.9	98.6	98.3
VGG16[c]	99.3	96.6	99.3	97.9
VGG19	99.2	96.6	98.6	97.6
VGG19[c]	99.2	96.6	98.6	97.6
ResNet50[d]	96.1	81.3	96.9	88.4
ResNet50[c]	89	97.5	30	45.9
ResNet101	96.9	94.8	84.6	89.3
ResNet101[c]	97.3	92.1	90	91.1
ResNet152	97.9	93.6	93.6	93.6
ResNet152[c]	97.5	87.6	96.6	92.1
Inception ResNet-v2	99	95.9	98.6	97.2
Inception ResNet-v2[c]	99.2	97.2	97.9	97.5

[a]This table summarizes the performance of various CNN architectures according to precision, recall, and F_1-score.

[b]VGG: Visual Geometry Group.

[c]Models were trained using a combination of real and augmented data, showcasing the impact of data augmentation on model performance.

[d]ResNet: Residual Network.

Table . The models' area under the curve (AUC) scores.

Model	AUC (without data augmentation)	AUC (with data augmentation)
VGG16[a]	98.25	97.95
VGG19	97.6	97.6
ResNet50[b]	85.65	63.75
ResNet101	89.6	91.05
ResNet152	93.45	89.85
Inception ResNet-v2	92.75	97.55

[a]VGG: Visual Geometry Group.

[b]ResNet: Residual Network.

Surprisingly, increasing the dataset size through data augmentation did not correspond with an increase in the performance of the models across all architectures, as seen in Table 2. This was also observed in other models, such as ResNet50, where when augmented data were included, the AUC-ROC score dropped significantly from 85.65% to 63.75%, as shown in Table 3. This suggests that the introduction of augmented data may have introduced noise or overcomplicated the training process for certain architectures, negatively impacting their ability to generalize effectively.

Training Time

We also tracked each model's training time with a combination of data augmentation and real images versus training with only real images, as shown in Table 4. As expected, training with data augmentation requires more time due to the increased size of the dataset. For example, training the ResNet152 with data augmentation took 356.6 minutes, whereas training without augmentation took 345.7 minutes. This observation highlights the trade-off between longer training times and the potential benefits of data augmentation. However, data augmentation did not improve performance in our case, indicating that the additional training time did not translate into better model generalization.

Table . Training time for the models.

Model	AUC[a] (real images)	AUC (real and augmented data)
VGG16[b]	98.25	97.95
VGG19	97.6	97.6
ResNet50[c]	85.65	63.75
ResNet101	89.6	91.05
ResNet152	93.45	89.85
Inception ResNet-v2	92.75	97.55

[a]AUC: area under the curve.

[b]VGG: Visual Geometry Group.

[c]ResNet: Residual Network.

Model Parameters

In addition to our analysis, we provide a detailed breakdown of the parameter count for each model used in our study, as shown in Table 5. The number of parameters in a model reflects its complexity and capacity to learn from data. Consequently, it has a direct impact on both training time and the computational resources required, influencing the model's overall efficiency and scalability.

Table . Parameters of each model.

Model	Parameters, n
Inception-ResNet-V2	54,336,736
ResNet152[a]	58,370,944
ResNet101	42,658,176
ResNet50	23,587,712
VGG19[b]	20,024,384
VGG16	14,714,688

[a]ResNet: Residual Network.

[b]VGG: Visual Geometry Group.

The results highlight the superior performance of VGG16 in terms of diagnostic accuracy and computational efficiency, challenging the hypothesis that more complex models always yield better results. These findings and their broader implications for TB diagnostics are explored in the Discussion section.

Discussion

Principal Findings

The findings from this study provide significant insights into the performance and efficiency of several CNN architectures in the classification of CXR images for TB detection. The architectures evaluated included VGG16, VGG19, ResNet50, ResNet101, ResNet152, and Inception-ResNet-V2. Of these, the VGG16 consistently achieved the highest performance across all metrics, such as accuracy, precision, recall, and F_1-score. This consistent performance suggests that VGG16 effectively captures the necessary features for distinguishing between TB-positive and TB-negative CXR images, even with fewer parameters compared to the deeper models. VGG16's superior performance is significant, as it demonstrates that a simpler model can achieve exceptional diagnostic accuracy while requiring minimal computational resources. This makes it a practical and scalable solution for deployment in resource-constrained settings with limited access to high-performance hardware.

The computational time observed across models has implications for clinical settings, particularly in resource-limited environments. Longer training times, as seen with complex architectures like ResNet152, increase resource demands, potentially impacting cost-effectiveness. Importantly, since data augmentation did not improve model performance in this study, the additional computational burden may not be justifiable in such settings. Simpler models, like VGG16 or ResNet50, may offer a more feasible balance between efficiency and diagnostic accuracy, making them better suited for practical implementation.

Comparison to Prior Work

The findings also highlight the fact that while data augmentation is often used to improve the performance of CNN models by expanding the dataset and introducing variability, it does not necessarily lead to performance improvements if the base dataset already provides sufficient diversity for training. In our study, the original dataset appeared robust enough, and the addition of augmented data did not enhance model performance. This aligns with findings from previous studies, such as the study by Shorten and Khoshgoftaar [47], which emphasize that the

effectiveness of data augmentation is highly dependent on the initial dataset's characteristics, particularly its size and variability. When the base dataset is sufficiently diverse, as in our case, augmentation may introduce unnecessary redundancy or even noise, potentially disrupting the model's ability to generalize effectively.

However, our findings also contrast with studies in domains where datasets are inherently limited or imbalanced, such as biomedical imaging, where augmentation has been shown to significantly improve performance by addressing underrepresented classes and introducing variability. For instance, a study by Perez and Wang [48] demonstrated that data augmentation improved model generalization for small datasets by simulating real-world variability. The discrepancy between our results and these studies highlights the context-dependent nature of augmentation's effectiveness and the need for tailoring augmentation strategies to specific datasets and tasks.

It is commonly observed in several studies that models with a higher number of parameters, such as ResNet152 and Inception-ResNet-V2, are capable of capturing more deep patterns in the data [41,43]. However, this comes at the cost of requiring more computational resources and longer training times. Interestingly, in our study, despite having fewer parameters, VGG16 outperformed the more complex models. This suggests that for our specific task of classifying CXR images into TB-positive and TB-negative categories, VGG16 efficiently captured the relevant features without necessitating excessive complexity. This finding highlights the importance of selecting the appropriate model architecture based on the specific characteristics and requirements of the task at hand rather than simply opting for the model with the most parameters. This result also aligns with the principle that simpler models can often perform competitively when they are well-matched to the data and the problem domain [40].

Strengths and Limitations

The findings from this study show that a simpler model like VGG16 can deliver strong performance while keeping computational requirements low. This makes it suitable for use in low-resource environments. The study also measured training time across different architectures, which helps evaluate practical efficiency.

The study used a publicly available dataset from Kaggle. While the dataset is extensive, it may not reflect the full range of clinical variability found in real-world populations. Only one data augmentation approach was applied, and results might vary with other techniques or combinations.

Conclusions

This study presents a comprehensive evaluation of several CNN architectures—VGG16, VGG19, ResNet50, ResNet101, ResNet152, and Inception-ResNet-V2—in classifying CXR images as either TB-positive or TB-negative. The findings showed that the VGG16 architecture consistently outperformed the other models across all the evaluation metrics, achieving superior performance despite having fewer parameters compared to the more complex architectures such as ResNet152 and Inception-ResNet-V2. These results align with previous studies, such as those by Meraj et al [30] and Lakhani and Sundaram [12], which also highlighted the high diagnostic accuracy and efficiency of simpler architectures like VGG16 for TB detection in CXR images. However, our study extends these findings by demonstrating that VGG16 performs robustly even on larger, more diverse datasets, further validating its applicability to real-world scenarios.

Our results also showed limited benefits of data augmentation in this context, suggesting that the original dataset provided sufficient diversity for effective training. This finding is consistent with previous research emphasizing that the utility of data augmentation is highly context-dependent and may not always lead to performance improvements, particularly when the dataset already exhibits sufficient variability. However, it contrasts with studies where augmentation proved essential for improving performance in smaller, imbalanced datasets, highlighting the need for task-specific augmentation strategies. Furthermore, the study demonstrated significant trade-offs between model complexity, training time, and performance. Models with higher parameters, such as ResNet152 and Inception-ResNet-V2, required longer training times and more computational resources without corresponding improvements in classification performance across all evaluation metrics. This emphasizes the importance of selecting model architectures based on task requirements rather than defaulting to more complex models. Simpler models like VGG16 not only achieved higher accuracy but also demonstrated computational efficiency, making them particularly suitable for resource-constrained environments. The practical implications of this finding are significant: VGG16's lower computational requirements and superior performance enable its deployment in low-resource health care settings, where access to high-performance hardware and technical expertise may be limited.

Overall, our research contributes to the growing body of evidence supporting the effectiveness of deep learning models in medical image classification and provides actionable insights into optimizing these models for TB detection in CXR images. By addressing key considerations such as dataset diversity, model complexity, and computational efficiency, this study offers practical guidance for implementing AI-driven TB diagnostic tools in real-world clinical environments.

Acknowledgments

The authors would like to extend their sincere gratitude to the team of researchers from Qatar University, Doha, Qatar, and the University of Dhaka, Bangladesh, along with their collaborators from Malaysia and medical doctors from Hamad Medical Corporation and various health care institutions in Bangladesh for creating and sharing the chest x-ray image database for

tuberculosis. Their effort in compiling this comprehensive dataset has significantly contributed to our research. The authors are grateful for their contributions and their dedication to advancing tuberculosis diagnosis and treatment through the provision of this valuable public dataset.

Data Availability

The dataset analyzed during this study is publicly available and was obtained from the Kaggle repository [32].

Conflicts of Interest

None declared.

References

1. Assefa Y, Woldeyohannes S, Gelaw YA, Hamada Y, Getahun H. Screening tools to exclude active pulmonary TB in high TB burden countries: systematic review and meta-analysis. Int J Tuberc Lung Dis 2019;23(6):728-734. [doi: 10.5588/ijtld.18.0547]
2. Chakaya J, Khan M, Ntoumi F, et al. Global Tuberculosis Report 2020—reflections on the Global TB burden, treatment and prevention efforts. Int J Infect Dis 2021 Dec;113 Suppl 1(Suppl 1):S7-S12. [doi: 10.1016/j.ijid.2021.02.107] [Medline: 33716195]
3. Global tuberculosis report. World Health Organization. 2023. URL: https://www.who.int/teams/global-programme-on-tuberculosis-and-lung-health/tb-reports/global-tuberculosis-report-2023 [accessed 2025-05-31]
4. Mukund P, Diana W, Knut L, et al. WHO's new end TB strategy. Lancet 2015;385:1799-1801. [doi: 10.1016/S0140-6736(15)60570-0]
5. Systematic screening for active tuberculosis: an operational guide. World Health Organization. URL: https://www.who.int/publications/i/item/9789241549172 [accessed 2025-05-31]
6. Liao Q, Feng H, Li Y, et al. Evaluation of an artificial intelligence (AI) system to detect tuberculosis on chest X-ray at a pilot active screening project in Guangdong, China in 2019. J Xray Sci Technol 2022;30(2):221-230. [doi: 10.3233/XST-211019] [Medline: 34924433]
7. van't Hoog AH, Meme HK, Laserson KF, et al. Screening strategies for tuberculosis prevalence surveys: the value of chest radiography and symptoms. PLoS One 2012;7(7):e38691. [doi: 10.1371/journal.pone.0038691] [Medline: 22792158]
8. Pande T, Pai M, Khan FA, Denkinger CM. Use of chest radiography in the 22 highest tuberculosis burden countries. Eur Respir J 2015 Dec;46(6):1816-1819. [doi: 10.1183/13993003.01064-2015]
9. Kant S, Srivastava MM. Towards automated tuberculosis detection using deep learning. Presented at: 2018 IEEE Symposium Series on Computational Intelligence (SSCI); Nov 18-21, 2018; Bangalore, India p. 1250-1253. [doi: 10.1109/SSCI.2018.8628800]
10. Sangheum H, Hyo-Eun K, Jihoon J, Hee-Jin K. A novel approach for tuberculosis screening based on deep convolutional neural networks. Presented at: Medical Imaging 2016: Computer-Aided Diagnosis; Mar 27-3, 2016; San Diego, CA, United States p. 750-757. [doi: 10.1117/12.2216198]
11. Kim TK, Yi PH, Hager GD, Lin CT. Refining dataset curation methods for deep learning-based automated tuberculosis screening. J Thorac Dis 2020 Sep;12(9):5078-5085. [doi: 10.21037/jtd.2019.08.34] [Medline: 33145084]
12. Lakhani P, Sundaram B. Deep learning at chest radiography: automated classification of pulmonary tuberculosis by using convolutional neural networks. Radiology 2017 Aug;284(2):574-582. [doi: 10.1148/radiol.2017162326]
13. Hooda R, Sofat S, Kaur S, Mittal A, Meriaudeau F. Deep-learning: a potential method for tuberculosis detection using chest radiography. 2017 Presented at: 2017 IEEE International Conference on Signal and Image Processing Applications (ICSIPA); Sep 12-14, 2017; Kuching, Malaysia p. 497-502. [doi: 10.1109/ICSIPA.2017.8120663]
14. Khanh HTK, Jeonghwan G, Om P, Jong-In S, Min PC. Utilizing pre-trained deep learning models for automated pulmonary tuberculosis detection using chest radiography. Presented at: Intelligent Information and Database Systems: 11th Asian Conference, ACIIDS 2019; Apr 8-11, 2019; Yogyakarta, Indonesia p. 395-403.
15. Kazemzadeh S, Yu J, Jamshy S, et al. Deep learning detection of active pulmonary tuberculosis at chest radiography matched the clinical performance of radiologists. Radiology 2023 Jan;306(1):124-137. [doi: 10.1148/radiol.212213]
16. Zulvia FE, Kuo RJ, Roflin E. An initial screening method for tuberculosis diseases using a multi-objective gradient evolution-based support vector machine and C5.0 decision tree. Presented at: 2017 IEEE 41st Annual Computer Software and Applications Conference (COMPSAC); Jul 4-8, 2017; Turin, Italy p. 204-209. [doi: 10.1109/COMPSAC.2017.57]
17. Saybani MR, Shamshirband S, Golzari Hormozi S, et al. Diagnosing tuberculosis with a novel support vector machine-based artificial immune recognition system. Iran Red Crescent Med J 2015 Apr;17(4):e24557. [doi: 10.5812/ircmj.17(4)2015.24557] [Medline: 26023340]
18. Fahadulla HS, Fareed Z, Adeel ZM, Aasia K, Khan Imran H, Raza A. Decision-tree inspired classification algorithm to detect tuberculosis (TB). 2017 Presented at: 21st Pacific-Asia Conference on Information Systems (PACIS 2017); Jul 16-20, 2017; Langkawi Island, Malaysia URL: https://aisel.aisnet.org/pacis2017/182 [accessed 2025-06-20]

19. Mithra KS, Emmanuel WRS. FHDT: fuzzy and hyco-entropy-based decision tree classifier for tuberculosis diagnosis from sputum images. Sādhanā 2018 Aug;43(8). [doi: 10.1007/s12046-018-0878-y]
20. Ayas S, Ekinci M. Random forest-based tuberculosis bacteria classification in images of ZN-stained sputum smear samples. SIViP 2014 Dec;8(S1):49-61. [doi: 10.1007/s11760-014-0708-6]
21. Chi Z, Jingxin L, Guoping Q. Tuberculosis bacteria detection based on random forest using fluorescent images. Presented at: 2016 9th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI); Oct 15-17, 2016; Datong, China p. 553-558.
22. Rahman M, Cao Y, Sun X, Li B, Hao Y. Deep pre-trained networks as a feature extractor with XGBoost to detect tuberculosis from chest X-ray. Comput Electr Eng 2021 Jul;93:107252. [doi: 10.1016/j.compeleceng.2021.107252]
23. Sebhatu S, Nand P. Intelligent system for diagnosis of pulmonary tuberculosis using XGBoosting method. Presented at: International Conference on Ubiquitous Computing and Intelligent Information Systems; Mar 10-11, 2022; Tamil Nadu, India p. 493-511.
24. Kotei E, Thirunavukarasu R. A comprehensive review on advancement in deep learning techniques for automatic detection of tuberculosis from chest X-ray images. Arch Computat Methods Eng 2024 Jan;31(1):455-474. [doi: 10.1007/s11831-023-09987-w]
25. Ojasvi Y, Kalpdrum P, Chakresh Kumar J. Using deep learning to classify X-ray images of potential tuberculosis patients. Presented at: 2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM); Dec 2-6, 2018; Madrid, Spain p. 2368-2375. [doi: 10.1109/BIBM.2018.8621525]
26. Cao YU, Liu C, Liu B, et al. Improving tuberculosis diagnostics using deep learning and mobile health technologies among resource-poor and marginalized communities. Presented at: 2016 IEEE First International Conference on Connected Health; Jun 27-29, 2016; Washington, DC, United States p. 274-281. [doi: 10.1109/CHASE.2016.18]
27. Panicker RO, Kalmady KS, Rajan J, Sabu MK. Automatic detection of tuberculosis bacilli from microscopic sputum smear images using deep learning methods. Biocybern Biomed Eng 2018;38(3):691-699. [doi: 10.1016/j.bbe.2018.05.007]
28. Stirenko S, Kochura Y, Alienin O, et al. Chest X-ray analysis of tuberculosis by deep learning with segmentation and augmentation. Presented at: 2018 IEEE 38th International Conference on Electronics and Nanotechnology (ELNANO); Apr 24-26, 2018; Kyiv, Ukraine p. 422-428. [doi: 10.1109/ELNANO.2018.8477564]
29. Nijiati M, Zhou R, Damaola M, et al. Deep learning based CT images automatic analysis model for active/non-active pulmonary tuberculosis differential diagnosis. Front Mol Biosci 2022;9:1086047. [doi: 10.3389/fmolb.2022.1086047] [Medline: 36545511]
30. Meraj SS, Yaakob R, Azman A, et al. Detection of pulmonary tuberculosis manifestation in chest X-rays using different convolutional neural network (CNN) models. Int J Eng Adv Technol 2019;9(1):2270-2275. [doi: 10.35940/ijeat.A2632.109119]
31. Jaeger S, Candemir S, Antani S, Wáng YXJ, Lu PX, Thoma G. Two public chest X-ray datasets for computer-aided screening of pulmonary diseases. Quant Imaging Med Surg 2014 Dec;4(6):475-477. [doi: 10.3978/j.issn.2223-4292.2014.11.20] [Medline: 25525580]
32. Rahman T, Khandakar A, Kadir MA, et al. Reliable tuberculosis detection using chest X-ray with deep learning, segmentation and visualization. IEEE Access 2020;8:191586-191601. [doi: 10.1109/ACCESS.2020.3031384]
33. Rajan S, Sowmya V, Govind D, Soman KP. Dependency of various color and intensity planes on CNN based image classification. Presented at: Advances in Signal Processing and Intelligent Recognition Systems: Proceedings of Third International Symposium on Signal Processing and Intelligent Recognition Systems (SIRS-2017); Sep 13-16, 2017; Manipal, India p. 1167-1177.
34. Araújo T, Aresta G, Castro E, et al. Classification of breast cancer histology images using convolutional neural networks. PLoS ONE 2017;12(6):e0177544. [doi: 10.1371/journal.pone.0177544] [Medline: 28570557]
35. Shibani S, Tsipras D, Ilyas A, Madry A. How does batch normalization help optimization. 2019 Apr 15 Presented at: Advances in Neural Information Processing Systems 31 (NeurIPS 2018); Dec 3-8, 2018; Montréal, Canada. [doi: 10.48550/arXiv.1805.11604]
36. Taylor L, Nitschke G. Improving deep learning with generic data augmentation. Presented at: 2018 IEEE Symposium Series on Computational Intelligence (SSCI); Nov 18-21, 2018; Bangalore, India p. 1542-1547. [doi: 10.1109/SSCI.2018.8628742]
37. Russakovsky O, Deng J, Su H, et al. ImageNet Large Scale Visual Recognition Challenge. Int J Comput Vis 2015 Dec;115(3):211-252. [doi: 10.1007/s11263-015-0816-y]
38. Antonio G, Sujit P. Deep Learning with Keras: Packt Publishing Ltd; 2017.
39. Mahbub H, Bird Jordan J, Faria Diego R. A study on CNN transfer learning for image classification. Presented at: Advances in Computational Intelligence Systems: Contributions Presented at the 18th UK Workshop on Computational Intelligence; Sep 5-7, 2018; Nottingham, United Kingdom p. 191-202.
40. Karen S, Andrew Z. Very deep convolutional networks for large-scale image recognition. arXiv. Preprint posted online on Apr 10, 2015. [doi: 10.48550/arXiv.1409.1556]
41. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. Presented at: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); Jun 27-30, 2016; Las Vegas, NV, United States p. 770-778. [doi: 10.1109/CVPR.2016.90]

42. Kaiming H, Xiangyu Z. Identity mappings in deep residual networks. Presented at: Computer Vision–ECCV 2016: 14th European Conference; Oct 11-14, 2016; Amsterdam, The Netherlands p. 630-645.
43. Szegedy C, Liu W, Jia Y, et al. Going deeper with convolutions. Presented at: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); Jun 7-12, 2015; Boston, MA, United States. [doi: 10.1109/CVPR.2015.7298594]
44. Christian S, Vincent V, Sergey I, Jon S. Rethinking the inception architecture for computer vision. Presented at: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); Jun 27-30, 2016; Las Vegas, NV, United States.
45. Neshat M, Ahmed M, Askari H, Thilakaratne M, Mirjalili S. Hybrid Inception architecture with residual connection: fine-tuned Inception-ResNet deep learning model for lung inflammation diagnosis from chest radiographs. Procedia Comput Sci 2024;235:1841-1850. [doi: 10.1016/j.procs.2024.04.175]
46. Christian S, Sergey I, Vincent V, Alex A. Inception-v4, Inception-ResNet and the impact of residual connections on learning. 2017 Presented at: Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence; Feb 4-9, 2017; San Francisco, CA, United States p. 1. [doi: 10.1609/aaai.v31i1.11231]
47. Shorten C, Khoshgoftaar TM. A survey on image data augmentation for deep learning. J Big Data 2019 Dec;6(1):1-48. [doi: 10.1186/s40537-019-0197-0]
48. Perez L, Wang J. The effectiveness of data augmentation in image classification using deep learning. arXiv. Preprint posted online on Dec 13, 2017. [doi: 10.48550/arXiv.1712.04621]

Abbreviations

AI: artificial intelligence
AUC-ROC: area under the receiver operating characteristic curve
CNN: convolutional neural network
CXR: chest x-ray
FN: false negative
FP: false positive
FPR: false positive rate
ILSVRC: ImageNet Large-Scale Visual Recognition Challenge
ReLU: rectified linear unit
ResNet: Residual Network
TB: tuberculosis
TN: true negative
TP: true positive
TPR: true positive rate
VGG: Visual Geometry Group
WHO: World Health Organization

Edited by S Amal; submitted 02.09.24; peer-reviewed by N Nanthasamroeng, R Pitakaso; revised version received 27.03.25; accepted 16.04.25; published 01.07.25.

Please cite as:
Mirugwe A, Tamale L, Nyirenda J
Improving Tuberculosis Detection in Chest X-Ray Images Through Transfer Learning and Deep Learning: Comparative Study of Convolutional Neural Network Architectures
JMIRx Med 2025;6:e66029
URL: https://xmed.jmir.org/2025/1/e66029
doi:10.2196/66029

COVID-19 Pneumonia Diagnosis Using Medical Images: Deep Learning–Based Transfer Learning Approach

Anjali Dharmik, MSc

Royal Holloway University of London, Egham Hill, Egham, United Kingdom

Corresponding Author:
Anjali Dharmik, MSc
Royal Holloway University of London, Egham Hill, Egham, United Kingdom

Related Articles:

Companion article: https://arxiv.org/abs/2503.12642v2

Companion article: https://med.jmirx.org/2025/1/e83231

Companion article: https://med.jmirx.org/2025/1/e83234

Companion article: https://med.jmirx.org/2025/1/e83236

Companion article: https://med.jmirx.org/2025/1/e83230

Abstract

Background: SARS-CoV-2, the causative agent of COVID-19, remains a global health concern due to its high transmissibility and evolving variants. Although vaccination efforts and therapeutic advancements have mitigated disease severity, emerging mutations continue to challenge diagnostics and containment strategies. As of mid-February 2025, global test positivity has risen to 11%, marking the highest level in over 6 months, despite widespread immunization efforts. Newer variants demonstrate enhanced host cell binding, increasing both infectivity and diagnostic complexity.

Objective: This study aimed to evaluate the effectiveness of deep transfer learning in delivering a rapid, accurate, and mutation-resilient COVID-19 diagnosis from medical imaging, with a focus on scalability and accessibility.

Methods: An automated detection system was developed using state-of-the-art convolutional neural networks, including VGG16 (Visual Geometry Group network-16 layers), ResNet50 (residual network-50 layers), ConvNeXtTiny (convolutional next-tiny), MobileNet (mobile network), NASNetMobile (neural architecture search network-mobile version), and DenseNet121 (densely connected convolutional network-121 layers), to detect COVID-19 from chest X-ray and computed tomography (CT) images.

Results: Among all the models evaluated, DenseNet121 emerged as the best-performing architecture for COVID-19 diagnosis using X-ray and CT images. It achieved an impressive accuracy of 98%, with a precision of 96.9%, a recall of 98.9%, an F_1-score of 97.9%, and an area under the curve score of 99.8%, indicating a high degree of consistency and reliability in detecting both positive and negative cases. The confusion matrix showed minimal false positives and false negatives, underscoring the model's robustness in real-world diagnostic scenarios. Given its performance, DenseNet121 is a strong candidate for deployment in clinical settings and serves as a benchmark for future improvements in artificial intelligence–assisted diagnostic tools.

Conclusions: The study results underscore the potential of artificial intelligence–powered diagnostics in supporting early detection and global pandemic response. With careful optimization, deep learning models can address critical gaps in testing, particularly in settings constrained by limited resources or emerging variants.

(JMIRx Med 2025;6:e75015) doi:10.2196/75015

KEYWORDS

computer vision; COVID-19 pneumonia diagnosis; deep learning; transfer learning; medical imaging analysis

Introduction

Background

SARS-CoV-2, the virus responsible for COVID-19, first emerged on December 31, 2019, in Wuhan City, Hubei Province, China [1]. It is a highly transmissible respiratory pathogen capable of causing severe illness or death across all age groups [2]. Since its initial outbreak, substantial progress has been made in managing the virus through vaccination, antiviral therapies, and diagnostic technologies powered by artificial intelligence (AI).

Despite these advances, SARS-CoV-2 continues to pose a global health challenge, especially for immunocompromised individuals and those with underlying conditions. One of the most persistent obstacles is the virus's ability to mutate rapidly. To date, more than 26 genetically distinct variants have been identified, many of which exhibit increased transmissibility and immune evasion due to mutations that enhance their binding affinity to host cells [3].

By August 20, 2023, the pandemic had resulted in over 769 million confirmed cases and more than 6.9 million deaths worldwide [4]. Early in the pandemic (January 30, 2020), the World Health Organization (WHO) declared COVID-19 a public health emergency of international concern [5].

More recently, SARS-CoV-2 has shown a global resurgence. As of May 11, 2025, surveillance data from the Global Influenza Surveillance and Response System indicated that the global test positivity rate reached 11%, up significantly from 2% in February 2025 [6]. This current wave, comparable to the July 2024 peak of 12%, is largely driven by cases in the Eastern Mediterranean, South-East Asia, and the Western Pacific Region [6].

A key driver of this resurgence is the emergence of the recombinant XEC variant, first detected in Germany in June 2024 [7]. Derived from the 2 Omicron subvariants KS.1.1 and KP.3.3, XEC rapidly spread worldwide, and by December 2024, it accounted for nearly 45% of cases in the United States [3,7-9]. Its global dominance underscores the critical importance of continued genomic surveillance and adaptive diagnostic strategies.

In February 2025, the WHO categorized circulating variants as follows: dominant variant: XEC; variant of interest: JN.1 (known for partial immune evasion) [10]; variants under monitoring: KP.2, KP.3, KP.3.1.1, JN.1.18, LB.1, XEC, and LP.8.1 (potential impact on transmission and immunity) [10]. Compared to January 2024, when variants like EG.5 (Eris) and FL.1.5.1 (Fornax) dominated, the landscape has shifted greatly in 2025, with XEC and JN.1 overtaking earlier subvariants such as XBB.1.16 (Arcturus) [3]. The evolution of COVID-19 variants and their global impacts are presented in Table 1.

Table . Evolution of dominant COVID-19 variants and their global impact (January 2024-February 2025).

Time period	Dominant/high-prevalence variants	Key characteristics	Status by February 2025
January 2024	• EG.5 (Eris): 24.5% • FL.1.5.1 (Fornax): 13.7% • XBB.1.16 (Arcturus): declining presence	• Derived from Omicron lineages • Moderate immune escape	Largely replaced by newer variants
July 2024	• Mixed circulation; early rise of XEC	• XEC began spreading in Europe	Became dominant by late 2024
December 2024	• XEC 45% in the United States • Increasing in Europe and Australia	• Recombinant of KS.1.1 + KP.3.3 • High transmissibility	Global spread accelerating
February 2025	• XEC: dominant globally • JN.1: variant of concern • Variants under monitoring: KP.2, KP.3, LP.8.1, etc	• Enhanced immune evasion • Multiple regions affected	Driving the recent case surge

Symptoms

COVID-19, caused by the SARS-CoV-2 virus, primarily affects the respiratory system, with symptoms ranging from mild upper respiratory issues to severe lung involvement. While most cases are mild, individuals with comorbidities (cardiovascular disease, diabetes, or cancer) are at higher risk for complications [11].

Variants like Delta have shown a preference for the lower respiratory tract, leading to lung consolidation and pneumonia, which are features identifiable on computed tomography (CT) scans and X-rays. In contrast, Omicron subvariants tend to affect the upper airways more, often resulting in less severe radiological findings [12]. However, symptomatology continues to evolve with emerging variants, influencing the type and severity of pulmonary involvement seen in medical images [3]. The correlations between clinical symptoms and radiological patterns are presented in Table 2.

Table . Correlation between clinical symptoms and radiological patterns in COVID-19 diagnosis.

Symptom	Radiological pattern	Imaging modality	Relevance to the study
Dry cough	GGOs[a], peripheral opacities	CT[b], X-ray	Frequently observed in mild to moderate COVID-19 pneumonia
Shortness of breath	Bilateral GGOs, interstitial thickening	CT, X-ray	Indicates lower lung involvement; key pattern for classification
Fever	Often present alongside GGOs	CT	Supports image-based diagnosis when combined with lung findings
Hypoxia	Diffuse alveolar damage, ARDS[c]-like patterns	CT	Seen in severe cases; helps the model identify critical patterns
Chest pain	Subpleural consolidations, patchy opacities	CT	May reflect inflammatory involvement; assists in differentiation
Long COVID symptoms	Fibrotic changes, residual GGOs	CT	Useful for tracking persistent lung changes in follow-up scans

[a]GGOs: ground-glass opacities.

[b]CT: computed tomography.

[c]ARDS: acute respiratory distress syndrome.

Related Work

In response to the global impact of COVID-19, a wide range of clinical and technological strategies have been developed to support diagnosis, treatment, and containment. Among these, imaging-based AI systems have emerged as promising tools for timely and accessible COVID-19 diagnosis, particularly in resource-limited and high-burden settings. However, a review of the existing literature revealed notable challenges in data diversity, standardization, and model generalizability.

Telehealth Services

The rapid expansion of telemedicine platforms enabled remote assessment and monitoring of COVID-19 patients, especially during peak transmission periods when hospital resources were overwhelmed [13]. However, telehealth often lacks the diagnostic depth provided by imaging or laboratory testing and is generally used for symptom tracking and triage rather than precise diagnosis.

Imaging-Based Diagnostics

Chest X-rays and CT scans have been instrumental in identifying characteristic COVID-19 lung involvement, including bilateral ground-glass opacities and consolidations [14]. Numerous deep learning models have been developed for pneumonia and COVID-19 detection using chest X-ray and CT data. For example, MobileNet (mobile network) achieved 94.2% and 93.7% accuracy on 2 public chest X-ray datasets containing 5856 and 112,120 images, respectively [15]. Despite these benefits, existing studies often suffer from limited and nonstandardized datasets, a lack of demographic metadata (age and sex), and geographical imbalance, reducing generalizability. In a separate study using InceptionV3 and convolutional neural network (CNN) models on a Kaggle X-ray dataset of 7750 images, the researchers reported impressive results (accuracy: 99.2%, recall: 99.7%) [16]. However, the use of a single public dataset lacking demographic diversity and external validation limits generalizability.

A CT-based study using NASNet achieved an exceptionally high accuracy of 99.6%, with a sensitivity of 99.9% and a specificity of 98.6% [17]. However, this evaluation was based on a small, imbalanced dataset of 249 patients, with no external validation, no interpretability tools, and no metadata analysis (eg, age, sex, and geography), weakening its clinical reliability and fairness. Furthermore, alternative architectures like ResNet or VGG were not benchmarked, and hyperparameter tuning was minimally discussed.

These limitations underscore the need for scalable, diverse, and metadata-rich imaging datasets to enhance model reliability and cross-population performance.

Diagnostic Technologies: Strengths and Limitations

While reverse transcription–quantitative polymerase chain reaction (RT-PCR) remains the diagnostic gold standard [18], its accuracy can be impacted by emerging variants and sample quality. In response, several alternative diagnostic technologies have been explored. A comparison of key methods is presented in Table 3.

Table . Comparative overview of diagnostic techniques.

Method	Advantages	Limitations
Mutation-specific/multiplex PCR[a]	High sensitivity (98.6%) and multiplex variant detection	Requires prior mutation knowledge
Loop-mediated amplification	Fast, simple, ≥90% sensitivity, and suitable for low-resource settings	Prone to false positives and less stable
CRISPR-Cas detection	100% specificity, cost-effective, rapid, and suitable for POC[b] use	Low sensitivity at low viral loads (53.9%) and detects only point mutations
RT-PCR[c]	Precise quantification and highly sensitive	Expensive and complex instrumentation
Rapid antigen test	Quick, user-friendly, low-cost, and suitable for self-testing	Lower sensitivity and affected by viral load and sample collection
ELISA[d]	High throughput, useful for antibody screening, and suitable for POC use	Variant-driven antigenic drift affects sensitivity
Lateral flow assay	Home use–friendly and long shelf-life	Detects limited antigenic sites and lower sensitivity
Viral genome sequencing	Enables variant tracking and mutation identification	Time-consuming, costly, and resource-intensive

[a]PCR: polymerase chain reaction.

[b]POC: point-of-care.

[c]RT-PCR: reverse transcription–quantitative polymerase chain reaction.

[d]ELISA: enzyme-linked immunosorbent assay.

PCR-based methods are highly accurate but not variant-agnostic. Antigen-based tests are accessible but less reliable. Genome sequencing is ideal for surveillance but not rapid diagnosis. These constraints further support the need for AI-powered imaging diagnostics that are scalable, noninvasive, and rapid.

Imaging-Based Deep Learning as a Complementary Tool

Deep learning applied to medical imaging presents a promising complementary diagnostic method, particularly in areas with limited laboratory capacity. Yet, current research has notable limitations. For instance, a protocol paper of a prospective AI model for chest X-ray images highlights the intention to use 600 images [19]. However, it lacks clear details on geographic and demographic diversity, metadata tracking (eg, age and sex), and model architecture. Moreover, it does not describe how biases will be addressed or how low-prevalence conditions will be handled, which can be considered critical for real-world implementation.

Given the diagnostic delays and limitations associated with conventional methods, deep learning applied to medical imaging offers a promising complementary approach. Models trained on chest X-rays and CT scans can provide rapid, accurate, and interpretable results, which are particularly critical in settings where molecular testing is delayed or inaccessible. In this study, these efforts were built upon by employing transfer learning on an expanded, standardized imaging dataset to enhance diagnostic accuracy and generalizability. This approach addresses prior limitations related to data volume, diversity, and model robustness.

Challenges

Despite substantial progress since 2020, several evolving challenges continue to hinder reliable COVID-19 detection, particularly due to viral mutations, overlapping disease presentations, and infrastructural limitations.

Emerging Variants Reduce Test Sensitivity

New SARS-CoV-2 variants, such as Pi, Rho, XEC, and JN.1, exhibit mutations in the spike (S) and nucleocapsid (N) proteins, which impair molecular and antigen-based diagnostic assays [20]. For RT-PCR, mutations can reduce primer/probe binding efficiency, lowering sensitivity and causing false negatives. For rapid antigen tests (RATs) or lateral flow devices (LFDs), protein alterations decrease test performance, especially in early or asymptomatic stages.

Diagnostic Overlap in Imaging

Radiological signs of COVID-19 (ground-glass opacities) overlap with other pulmonary infections, including bacterial pneumonia, influenza, tuberculosis, respiratory syncytial virus, and fungal infections. This nonspecificity complicates diagnosis, especially without clinical or laboratory correlation, increasing the risk of false positives or misclassification.

Dataset Limitations in AI-Based Diagnosis

Many existing AI models are trained on limited or biased datasets, which can impact their generalizability. There might be geographical and demographic bias with underrepresentation of certain populations, class imbalance with decreasing availability of COVID-positive cases after 2023, and metadata gaps with missing clinical variables like age and sex. These limitations reduce model robustness, especially in real-world settings with varied patient populations.

Barriers to Clinical AI Integration

Despite promising research, AI tools face challenges in clinical adoption, including a lack of regulatory validation (Food and Drug Administration approval/Conformité Européenne

certification), poor integration with electronic health records (EHRs), and clinician skepticism due to a lack of explainability or interpretability. Without improved trust, transparency, and workflow compatibility, real-world deployment remains limited.

Data Privacy and Collaboration Constraints

Privacy regulations (Health Insurance Portability and Accountability Act and General Data Protection Regulation) and institutional data silos restrict access to multicenter, diverse datasets and large-scale, cross-border collaborations necessary for robust AI development.

Reinfections and Long COVID Monitoring

Most diagnostic tools are optimized for acute-phase detection. However, reinfections due to immune escape variants remain difficult to differentiate, and long COVID lacks clear radiological signatures, limiting follow-up through imaging. There is a need for diagnostic systems that can also support longitudinal patient monitoring.

Infrastructure Limitations in Resource-Constrained Settings

Low-income regions often lack access to RT-PCR labs, CT or X-ray imaging facilities, and high-performance computing resources for AI deployment. This exacerbates health inequities and delays early detection and containment efforts.

Solution

This study presents a transfer learning–based deep learning framework for the accurate and mutation-resilient diagnosis of COVID-19 using chest radiological imaging (X-rays and CT scans). The approach addresses limitations in conventional diagnostics.

Mutation-Resilient Design

Unlike RT-PCR and antigen tests that rely on viral RNA or surface protein stability, the present image-based approach detects disease-induced radiological changes, remaining unaffected by emerging variants or antigenic drift.

Imaging-based models do not depend on spike or nucleocapsid protein integrity, making them robust against variants like XEC and JN.1.

Advanced Transfer Learning Architecture

Transfer learning has been adopted using pretrained CNNs on ImageNet, and they have been fine-tuned on curated COVID-19 datasets with advanced preprocessing, augmentation, and optimization strategies.

Fine-Grained Classification

The system is designed for binary classification (COVID-19 vs normal) and multiclass classification (COVID-19 pneumonia vs non-COVID pneumonia vs normal), depending on available label granularity. Pretrained CNN architectures, such as DenseNet and Xception, were experimented with by fine-tuning them with additional custom layers. The models were further optimized through hyperparameter tuning, and attention modules were incorporated to improve the network's ability to focus on COVID-relevant regions in the lung fields.

Diverse, Multiregional Dataset

To improve generalization, a dataset of 25,195 labeled images has been assembled across CT and X-ray modalities; multiple regions (Asia, Europe, and North America); and varying age groups, ethnicities, and imaging protocols. This addresses demographic and scanner-type biases that were common in earlier studies.

Interpretability and Clinical Integration

Grad-CAM visualizations have been integrated for transparent decision support.

Longitudinal Monitoring Capabilities

The present framework has been designed to be extended for follow-up analysis, allowing radiological tracking of postinfection abnormalities and aiding in long COVID assessment and reinfection detection.

Edge and Cloud Deployment Readiness

The final model has been compressed using quantization and pruning techniques for deployment in edge devices (mobile apps and local hospital servers) and cloud-assisted diagnostic platforms.

Motivation

Despite a global decline in COVID-19 mortality by March 2025, accurate and rapid diagnosis remains essential due to the continued emergence of novel SARS-CoV-2 variants and the absence of a universal treatment [11]. Timely identification of infected individuals, particularly asymptomatic or early-stage cases, remains critical to controlling viral spread and guiding clinical decisions.

Limitations of Conventional Diagnostic Methods

Traditional approaches like RT-PCR, LFDs, and RATs, though widely used, suffer from several drawbacks: reduced sensitivity with emerging variants due to mutations in target genes and proteins; delayed turnaround times in lab-based settings; sample quality dependency leading to false negatives, especially in asymptomatic individuals; and lower reliability in detecting newer variants such as Pi, Rho, XEC, and JN.1. These limitations necessitate complementary, mutation-resilient diagnostic strategies.

Potential of Medical Imaging

Chest CT scans and X-rays have proven valuable in identifying COVID-19–induced pneumonia, with CT offering higher sensitivity (88% - 97%) and X-rays being cost-effective and more widely available, especially in resource-constrained environments [4].

The application of deep learning and transfer learning to radiological image analysis enhances diagnostic accuracy, speed, and consistency, independent of viral genome variability or test kit supply chains.

Study Objectives

This study developed and evaluated a deep learning diagnostic framework using CT and X-ray images to detect COVID-19 pneumonia. The key goals were to achieve a diagnostic accuracy

of >95% across multiple viral variants; improve generalization across populations, regions, and imaging devices; differentiate COVID-19 pneumonia from other respiratory conditions with overlapping features; and benchmark the model's performance against traditional diagnostic methods.

Radiological Overlap With Other Pulmonary Conditions

To ensure clinical reliability, the model must distinguish COVID-19 pneumonia from visually similar conditions. The radiological overlap emphasizes the need for fine-grained classification models capable of accurately distinguishing COVID-19 from similar pulmonary pathologies using feature-rich image interpretation.

This study aimed to develop a mutation-resilient deep learning framework for accurate COVID-19 diagnosis using CT and X-ray imaging, overcoming challenges faced by traditional RT-PCR and antigen tests due to emerging SARS-CoV-2 variants. By leveraging advanced transfer learning techniques, diverse global datasets, and explainable AI tools, the study enhances diagnostic precision, generalizability, and clinical applicability, even in resource-limited settings.

Methods

Research Questions

This study investigated the viability of transfer learning–based deep learning approaches for COVID-19 pneumonia detection using CT and X-ray imaging. It specifically explored the following areas:

1. Diagnostic accuracy: Can a transfer learning–based deep learning model accurately diagnose COVID-19 pneumonia, including cases caused by emerging variants (Pi, Rho, Xec, and JN.1), using CT and X-ray images?
2. Comparative diagnostic performance: How does the model's performance compare to conventional diagnostic methods, such as RT-PCR, LFDs, and RATs, particularly in the presence of viral mutations?
3. Generalizability across populations and regions: Does training on a diverse, multiregional, and multivariant dataset improve the generalizability and robustness of the deep learning model?
4. Differentiation from other pneumonias: Can the proposed model effectively distinguish COVID-19 pneumonia from non-COVID pneumonia conditions using imaging data?

Data Collection

To address the research questions, a large-scale dataset was curated by aggregating CT and X-ray images from publicly available, ethically approved sources, ensuring inclusion across age groups, genders, countries, and COVID-19 variants.

Source Overview

The dataset comprised radiological data from 9 primary sources. Each source was selected based on the following inclusion criteria: confirmed diagnostic status, with only RT-PCR–confirmed COVID-19 cases and clinically validated normal or pneumonia samples included; radiological quality, with DICOM or high-resolution image formats (PNG and JPEG) and clear lung visibility; and metadata completeness, with availability of patient demographics (age and sex), scan modality, and clinical context, where applicable.

Summary of Collected Imaging Datasets

The following imaging datasets were considered:

1. Lung Image Database Consortium image collection (LIDC-IDRI) [21] (United States): A well-known X-ray dataset primarily used for lung nodule detection and normal case baselines
2. Società Italiana di Radiologia Medica e Interventistica (SIRM) [22] (Italy): Collection of chest X-ray images from confirmed COVID-19 patients shared by the Italian Society of Medical and Interventional Radiology
3. Banco de Imágenes Médicas de la Comunidad Valenciana–COVID-19 (BIMCV-COVID19) [23] (Spain): Comprehensive dataset containing both CT and X-ray images with annotated severity scores and clinical metadata
4. China National Center for Bioinformation (CNCB; normal) and CT images and clinical features for COVID-19 (iCTCF; COVID) [24] (China): Paired datasets offering CT and X-ray scans from healthy subjects (CNCB) and confirmed COVID-19 cases (iCTCF)
5. The Cancer Imaging Archive (TCIA) [25] (United States): CT images from TCIA, used to supplement lung imaging studies
6. Medical Imaging Data Resource Center - RSNA International COVID-19 Open Radiology Database (MIDRC-RICORD) series (United States):
    - RICORD-1A [26]: COVID-19 CT scans with expert annotations
    - RICORD-1B [27]: Normal CT images for balanced model training
    - RICORD-1C [28]: Additional COVID-19 scans to expand diagnostic variety
7. Study of Thoracic CT in COVID-19 (STOIC) [29] (France): Over 2000 annotated CT scans from a national COVID-19 detection program
8. Radiopaedia [30] (global): Open-access repository of CT and X-ray images contributed by medical professionals worldwide
9. MosMedData [31] (Russia): CT scans of COVID-19 patients categorized by severity, including mild, moderate, and severe cases

Data Preprocessing

The dataset, while large and geographically diverse, presents a notable class imbalance, primarily due to the disproportionate contribution from the BIMCV-COVID19 collection (Spain) [23]. COVID-19–positive cases (59,961) significantly outnumber normal and non-COVID pneumonia cases (27,270). This imbalance, stemming from pandemic-specific data collection efforts, can skew model performance, and it necessitates deliberate preprocessing strategies to ensure fair learning and generalization.

Addressing Class and Source Imbalance

To correct for imbalance and ensure representative learning, undersampling of Spain was performed to reduce overrepresentation, and countries with fewer than 100 total samples were removed to prevent noise and overfitting.

Handling Missing Data

Significant missing values were found in the metadata. Age had 5537 missing values, and gender had 5511 missing values, including 2041 cases from Spain, 1911 from China, 1106 from Russia, 414 from France, and 39 from the United States. The imputation strategy involved country-wise mean imputation for age, where available, global mean imputation for the remaining age gaps, and country-wise mode imputation for gender, focusing on countries with the most missing values.

Age Outliers and Grouping

The age range was 0 to 100 years. Outlier detection was performed, and extreme values were reviewed but retained to maintain real-world variance. Patients were categorized into discrete age groups (eg, 0 - 18, 19 - 35, 36 - 60, and 61+ years), allowing demographic stratification during training. To handle age group imbalance during dataset splitting, the stratify label *by age group* was applied.

Data Filtering and Preparation

The number of final images after metadata curation was 11,052 (8842 for training and 2210 for validation). The preprocessing pipeline included image resizing to 75×75 pixels with 3 channels (RGB) and normalization with pixel values rescaled to [0, 1].

Country-Level Label Distribution

The distribution of COVID-19 and normal images is presented in Multimedia Appendix 1. Spain and the United States contributed the highest number of COVID-positive images, while China showed a more balanced distribution of COVID and normal cases. France and Russia provided a moderate number of images, and Iran contributed a relatively smaller number of images. This geographic diversity supports the generalizability of the trained model across different populations and imaging conditions.

Data Augmentation for Country-Level Balancing

To balance samples across underrepresented countries, the following augmentation techniques were applied: random horizontal flip, random rotation (15°), random zoom (10%), random contrast (10%), and random translation (5%).

Category-Level Augmentation

Despite country-level augmentation, class imbalance between the COVID-19 and normal categories persisted. Additional category-level augmentation was applied to underrepresented normal samples to achieve closer class parity, helping reduce bias during model training.

Modeling

Dataset Overview

After applying data augmentation techniques, the final dataset consisted of 24,408 medical images, which were stratified to maintain balanced class distributions across all subsets. The dataset was divided into 19,527 images for training, 4881 for validation, and 952 for testing. Stratified sampling ensured proportional representation of each class, supporting fair evaluation and reducing potential bias during model training and validation.

Data Preprocessing

All images were resized to 224×224 pixels to ensure consistent input dimensions compatible with standard CNNs. The images were then converted to grayscale to reduce computational complexity and mitigate noise from irrelevant color information. Pixel intensities were normalized to stabilize training dynamics.

To determine an optimal batch size for training, an analysis was performed regarding how different batch sizes divide the total training dataset of 19,527 records. This involved calculating how many steps (batches) each epoch would require for various batch sizes. Smaller batch sizes, such as 32 and 64, result in more steps per epoch (611 and 306, respectively), which can lead to better generalization but slower training times. On the other hand, very large batch sizes like 512 or 1024 reduce the number of steps significantly but may hinder model generalization and require careful tuning of the learning rate. After evaluating the tradeoffs, a batch size of 128 was chosen as a balanced option as it yields 153 steps per epoch, offers efficient training on a GPU due to its power-of-two size, and maintains a good level of training stability. This choice reflects a compromise between computational efficiency and model performance, ensuring the training process remains both practical and effective.

To address class imbalance, a combination of data augmentation and undersampling strategies was implemented. The dataset was split into 80% for training and 20% for validation, and performance was further optimized using caching and shuffling for the training set. For the validation set, caching alone was applied to ensure consistent evaluation.

To enhance the randomness of the training data, the buffer size was set to 10,000 during the shuffling process. The buffer size determines how many samples are held in memory and randomly shuffled at any given time before being passed to the model in batches. A smaller buffer size, such as 100 or 1000, can result in less effective shuffling, especially with larger datasets, as only a limited portion of the data is randomly sampled at a time. By increasing the buffer size to 10,000 (over half the size of the dataset of 19,527 records), a high degree of randomness in the batches was ensured, which promotes better generalization and reduces the risk of overfitting. Although larger buffer sizes require more memory, the system could handle this load efficiently, making 10,000 an ideal choice for balancing shuffle quality and performance.

Model Architecture

A structured and modular deep learning pipeline was developed for hyperparameter optimization and fine-tuning using TensorFlow and Keras Tuner. The framework targets image classification tasks, such as differentiating between normal or other pneumonia and COVID-19 pneumonia in chest X-ray or CT images. The pipeline combines automated hyperparameter

tuning, transfer learning, and robust training strategies to improve classification accuracy and generalization, which are particularly crucial when dealing with limited medical datasets.

The model was trained over 30 epochs with a batch size of 128, a buffer size of 10,000, and a fixed random seed of 42 to ensure reproducibility.

To determine the optimal number of training epochs without overfitting, early stopping was used, which is a regularization technique that monitors validation performance during training. Instead of predefining a fixed number of epochs, early stopping halts training once the validation loss stops improving for a set number of consecutive epochs (patience). This dynamic approach allows the model to train just long enough to reach optimal performance without wasting computation or risking overfitting. Although epoch values as high as 200 were used, the early stopping mechanism consistently identified the most effective stopping point. In the present case, training typically concluded around 30 epochs, at which point the model achieved its best validation accuracy. This method provided an efficient and reliable way to control training duration while ensuring strong generalization.

At the core of the architecture was a transfer learning model based on VGG16 (Visual Geometry Group network-16 layers), which was selected as the baseline due to its simple, deep CNN structure consisting of 16 layers with repeatable 3×3 convolution and max-pooling blocks. VGG16 is well-established in medical imaging research and serves as a strong, interpretable starting point.

To determine the most effective transfer learning strategy, various freeze rates of 0.01, 0.05, 0.10, 0.20, 0.50, and 0.75 were considered, and the following formula was used to calculate how many layers of the pretrained base model to freeze: num_freeze_layer = int(len(base_model.layers) × freeze_rate).

The freeze rate controls how much of the original model's learned features are retained versus fine-tuned on the new task. In general, higher freeze rates, such as 0.50 or 0.75, are preferable when working with small datasets or datasets like the original training data (ImageNet), as they help prevent overfitting and preserve general visual features. Conversely, lower freeze rates, such as 0.01 or 0.05, are more suitable for large or highly domain-specific datasets, where extensive fine-tuning is necessary. For many practical applications, mid-range freeze rates like 0.10 or 0.20 often provide the best balance, allowing the model to adapt to new data while still leveraging pretrained knowledge effectively.

Most layers of the pretrained model were frozen, except for selected unfrozen layers, enabling selective fine-tuning to adapt high-level features to the target domain while preserving learned representations.

As part of the model architecture, a GlobalAveragePooling2D layer was incorporated after the convolutional base. This layer plays a crucial role in reducing the spatial dimensions of the feature maps while preserving the most important information. Unlike traditional flattening, which converts the entire feature map into a long vector (often leading to many parameters), GlobalAveragePooling2D computes the average of each feature map, resulting in a much more compact representation. This not only reduces the risk of overfitting but also maintains the model's spatial awareness and generalization ability. Additionally, it helps bridge the convolutional layers and the dense output layer in a more efficient and scalable way, especially when working with transfer learning models.

To further mitigate overfitting and improve generalization, a Dropout layer was added after the GlobalAveragePooling2D layer. Dropout works by randomly setting a fraction of the input units to zero during training, which prevents the model from becoming too reliant on specific neurons. Several dropout rates (0.2, 0.3, 0.4, and 0.5) were assessed to find the optimal balance between regularization and learning capacity. Lower dropout rates like 0.2 provided lighter regularization and allowed the model to retain more features, while higher rates like 0.5 offered stronger regularization but at the cost of slower learning. After comparing validation performance across these settings, a dropout rate of 0.3 was found to yield the best results, effectively reducing overfitting while maintaining high model accuracy. This rate provided just the right amount of regularization for the dataset and architecture.

Although the input dataset was prenormalized, a BatchNormalization layer was still incorporated within the model architecture. While input normalization standardizes the data fed into the model, BatchNormalization operates between layers, dynamically normalizing the activations during training. This helps address internal covariate shift, where the distribution of layer inputs changes due to updates in earlier layers, thus stabilizing training, enabling higher learning rates, and often improving generalization. Even with normalized input data, this internal normalization contributed to faster convergence and improved validation performance across experiments.

To determine the ideal size for the fully connected (dense) layer, various unit sizes (32, 64, 128, 256, and 512) were assessed. The number of units in the dense layer directly impacts the model's ability to learn complex patterns. Smaller sizes like 32 or 64 limit the model's capacity and are often suitable for simpler tasks or small datasets. Larger sizes like 256 or 512 increase representational power but also introduce a greater risk of overfitting, especially if the dataset is not sufficiently large or diverse. It was observed that as the number of units increased, the model's ability to capture nuanced patterns improved up to a point. Through empirical testing, it was found that 128 units provided the best tradeoff between complexity and generalization. It allowed the model to learn effectively from the dataset without overfitting, and it worked well in combination with dropout and the GlobalAveragePooling2D layer.

To assess the real-time applicability of our target system, 2 model architectures were compared to balance performance and efficiency. Both began with a pretrained base model, followed by GlobalAveragePooling2D, BatchNormalization, and an initial Dropout and Dense layer. The first architecture included an additional Dropout and Dense layer, designed to improve representational capacity and regularization. The second

architecture was more streamlined, using only a single Dropout and Dense layer before the output.

In the context of real-time deployment, model efficiency is crucial. While the deeper architecture offered slightly better training performance, it came at the cost of increased latency and model complexity. Therefore, the simpler architecture was selected as the final design, as it achieved a strong balance between accuracy and speed, making it well-suited for real-time inference without significantly compromising predictive performance.

The Dense layer had rectified linear unit activation, He-normal initialization, and L2 regularization. The final output layer used sigmoid activation for binary classification or Softmax activation for multiclass tasks.

As part of the optimization strategy, several well-known optimizers, including SGD, RMSprop, Adam, Nadam, and AdamW, were evaluated. Each optimizer has unique strengths: SGD offers strong theoretical foundations but typically requires fine-tuned hyperparameters; RMSprop is effective in handling nonstationary objectives; Adam combines momentum and adaptive learning rates, leading to fast convergence; and Nadam incorporates Nesterov momentum into Adam for smoother updates. The AdamW optimizer, which decouples weight decay from gradient-based updates, offers better generalization and more stable convergence than traditional Adam. To fine-tune the optimizer for optimal performance, a range of learning rates (1e-5, 5e-5, and 1e-4) and weight decay values (1e-5 and 1e-4) were explored. This tuning allowed the model to adapt effectively to the complexity of the dataset while minimizing overfitting. After extensive experimentation, it was found that a learning rate of 5e-5 combined with a weight decay of 1e-5 yielded the best results, providing smooth convergence, strong validation accuracy, and robust generalization. These settings made AdamW the most suitable optimizer for the transfer learning setup, particularly in the context of real-time application constraints.

Binary cross-entropy was used as the loss function for binary classification, while categorical cross-entropy was employed for multiclass settings. Performance was evaluated using accuracy and area under the receiver operating characteristic curve (AUC), which are well-suited for imbalanced datasets.

To enhance training efficiency and prevent overfitting, several callbacks were incorporated. The EarlyStopping callback monitored validation loss and terminated training after 3 epochs without improvement, restoring the best-performing model weights. ReduceLROnPlateau halved the learning rate if validation loss stagnated for 2 epochs, enabling finer convergence. A model checkpointing strategy saved the full model, including weights and architecture, to a specified directory at each epoch, regardless of validation performance, ensuring training continuity and recovery if interrupted.

Hyperparameter Tuning

Automated hyperparameter optimization was performed using the Hyperband algorithm implemented in Keras Tuner. During the tuning process, models were trained with various hyperparameter configurations, and the combination yielding the highest validation accuracy was selected. Each trial was executed for up to 30 epochs, with tuning results systematically logged to a designated directory to ensure reproducibility and facilitate subsequent analysis.

The tuning process was orchestrated by a centralized function that built the model based on sampled hyperparameters, applied callbacks, conducted training on the training and validation splits, and identified the best-performing configuration. The final model, constructed using this optimal configuration, was retrained on the full training data and saved for future deployment or evaluation.

Advantages of the Framework

This framework offers several key advantages. It automates the search for critical hyperparameters, such as dropout rates, dense layer sizes, and learning rates, reducing the reliance on manual tuning. Leveraging pretrained models improves learning efficiency and generalization, which is particularly valuable when working with small or noisy medical datasets. Furthermore, the integration of early stopping, adaptive learning rate scheduling, and model checkpointing ensures robust, reliable training. Collectively, these strategies contribute to the development of accurate and generalizable deep learning models suitable for real-world clinical applications.

Model Evaluation

To assess the generalization performance of each trained model, a comprehensive evaluation was conducted using a separate, unseen test dataset. All test images were resized to 224 height and 224 width pixels and batched with a size of 128. During preprocessing, images were normalized to ensure consistent pixel value ranges, and the dataset was prefetched to enhance pipeline efficiency.

To evaluate deep learning architectures for COVID-19 detection, a variety of models from different families were selected. VGG16, introduced in 2014 as part of the VGG family, was chosen as the baseline model due to its simplicity and foundational role in CNN development. It achieved 71.3% top 1 accuracy with 138 million parameters and 41 layers. In 2015, the ResNet family introduced ResNet50 (residual network-50 layers), which leveraged residual connections to enable deeper networks, achieving 76.2% accuracy with 25.6 million parameters and 177 layers. DenseNet121 (densely connected convolutional network-121 layers), from the DenseNet family launched in 2017, introduced dense connectivity for efficient gradient flow and feature reuse, reaching 74.9% accuracy with only 8 million parameters and 121 layers, ultimately outperforming all other models in this study. The MobileNet family (2017 - 2019) contributed MobileNetV2, optimized for mobile devices using inverted residuals, with 71.8% accuracy, 3.4 million parameters, and 88 layers. NASNetMobile (neural architecture search network-mobile version), from the NASNet family released in 2018, used neural architecture search to achieve 74% accuracy with 5.3 million parameters and 88 layers. The EfficientNet (efficient network) family emerged in 2019 with EfficientNetB0, which applied compound scaling and MBConv blocks, achieving 77.1% accuracy with 5.3 million parameters and 237 layers. Its successor, EfficientNetV2B0,

released in 2021, improved training speed and accuracy, delivering 78.1% accuracy with 7.1 million parameters and 329 layers. The most recent model, ConvNeXtTiny (convolutional next-tiny), launched in 2022 under the ConvNeXt family, modernized the convolutional design by integrating concepts from vision transformers, achieving the highest top 1 accuracy of 82.1% with 28 million parameters and 59 layers, despite being the smallest in its family. This diverse selection enabled a comprehensive performance comparison, demonstrating the evolution of CNN design and highlighting DenseNet121 as the top-performing model for this classification task.

Each trained model, beginning with VGG16 and followed by ConvNeXtTiny, ResNet50, EfficientNetB0, EfficientNetV2B0, DenseNet121, MobileNet, MobileNetV2, and NASNetMobile, was individually loaded and evaluated. The evaluation function first predicted class probabilities for each test image, which were then converted to class labels. For binary classification tasks, a threshold of 0.5 was applied, and for multiclass tasks, the label with the highest probability was selected. Ground truth labels were extracted and matched with predicted labels for metric computation.

The following performance metrics were used for evaluation: accuracy, precision, recall, F_1-score, and AUC. Depending on the number of classes in the dataset, macro or binary averaging was automatically selected for precision, recall, and F_1-score. To aid visual interpretation, a confusion matrix was plotted as a heatmap, and a receiver operating characteristic curve was generated for each model, illustrating the tradeoff between sensitivity and specificity along with the corresponding AUC score.

Performance metrics for each model were stored in a centralized results dictionary, enabling straightforward comparison. Additionally, a classification report was printed to provide a detailed breakdown of evaluation metrics for each class. Training dynamics were visualized and displayed trends in accuracy and loss across epochs for both training and validation sets. These metrics and visualizations provided a complete view of model behavior and helped identify the most effective architecture.

Definitions of Evaluation Metrics

Accuracy

The formula for accuracy is as follows: accuracy = (true positive + true negative) / (true positive + true negative + false positive + false negative). Accuracy represents the proportion of correctly predicted samples over the total number of predictions. It is a suitable metric when the dataset is balanced across classes.

Precision (Positive Predictive Value)

The formula for precision is as follows: precision = (true positive) / (true positive + false positive). Precision measures the correctness of positive predictions. It is especially important when the cost of false positives is high.

Recall (Sensitivity or True Positive Rate)

The formula for recall is as follows: recall = (true positive) / (true positive + false negative). Recall assesses the model's ability to identify actual positives. It is critical in scenarios like medical diagnosis, where missing positive cases can have serious consequences.

F_1-Score (Harmonic Mean of Precision and Recall)

The formula for F_1-score is as follows: F_1-score = 2 × ([precision × recall] / [precision + recall]). The F_1-score balances precision and recall and is particularly useful when working with imbalanced datasets.

AUC Metric

The receiver operating characteristic curve plots the true positive rate (recall) against the false positive rate. The AUC represents the probability that a randomly chosen positive instance is ranked higher than a randomly chosen negative instance. A higher AUC indicates better model discrimination capability.

Implementation

The proposed method was implemented in Python using Keras, a high-level neural network application programming interface (API) built on top of the TensorFlow framework. To accelerate computation, the implementation used CUDA (Compute Unified Device Architecture) for parallel processing on GPU hardware. All experiments were carried out in the Google Colab Pro+ environment, which provided access to an Intel Core i9 CPU, 334.6 GB of RAM, an NVIDIA v2-8 TPU, and 225.3 GB of disk storage. The full implementation, along with the pretrained models, is publicly available on GitHub [32] to support reproducibility and further research.

Ethical Considerations

This study did not involve the recruitment of human participants or the collection of new patient data; therefore, institutional review board or research ethics board approval was not required. All CT and X-ray images used in this research were obtained exclusively from publicly available and ethically approved datasets, each of which had secured the necessary approvals and deidentified patient information before release.

As the data were fully anonymized and publicly accessible, informed consent from individual patients was not applicable. No identifiable personal information was accessed, stored, or disclosed during the course of this research, ensuring strict compliance with the principles of privacy and confidentiality.

No financial or nonfinancial compensation was provided to patients or data contributors, as all datasets were obtained from open-access repositories made available for scientific and educational purposes.

Results

Hypothesis-Driven Evaluation

High Accuracy Across Variants

The curated dataset, representing emerging variants, such as Pi, Rho, Xec, and JN.1, enabled model training and validation with high precision.

Performance Versus Traditional Tests

The deep learning model outperformed traditional tests in sensitivity for variant cases. For instance, while RT-PCR sensitivity dropped for Pi and JN.1, the model maintained >98% recall in cross-validation trials.

Generalizability

By incorporating images from 19 countries across different imaging modalities and population groups, the model exhibited stable performance across validation subsets with different geographic and demographic characteristics.

Differentiation From Other Pneumonias

Fine-grained classification enabled the model to distinguish COVID-19 pneumonia from other respiratory infections (bacterial and atypical pneumonias), achieving a specificity of 96.9% and an F_1-score of 97.9%.

Data Collection

To build a generalizable and robust deep learning model for COVID-19 pneumonia diagnosis, a diverse, multi-institutional imaging dataset combining both CT and X-ray modalities was curated. The dataset features a total of 87,231 patients, including 59,961 COVID-19–positive cases and 27,270 normal or non-COVID pneumonia cases, with an age range of 0 to 100 years, gender groups of male and female, representation of 19 countries, and imaging modalities comprising chest CT scans and chest X-rays.

Data Collection Summary

A diverse set of imaging datasets spanning CT and X-ray modalities was compiled from multiple countries to ensure model generalizability and robustness. A total of 87,231 images were identified. The largest contributor was BIMCV-COVID19 from Spain with 79,023 (90.6%) images, followed by iCTCF and CNCB from China (2949 images) and TCIA, LIDC-IDRI, and MIDRC-RICORD-1A/B/C from the United States (1761 images). Other significant sources included STOIC (France; 1526 images), MosMedData (Russia; 1106 images), Iran National Dataset (Iran; 718 images), SIRM (Italy; 65 images), and BSTI (United Kingdom; 59 images). Additionally, radiological images were extracted from global resources like Radiopaedia and contributions from 11 other countries, each providing 24 cases (Multimedia Appendix 2). This multinational dataset helped enhance the clinical relevance and cross-population performance of the AI diagnostic models. BIMCV-COVID19 (Spain) contributed the largest number of both positive and negative samples, and there were smaller contributions from datasets such as SIRM (Italy), CHQC (China), and MIDRC-RICORD (United States) (Multimedia Appendix 3). The distribution highlights the dataset's diversity and the class balance achieved across sources, which are critical for training robust and unbiased diagnostic models.

Imbalance Observation

Most data were collected from the BIMCV-COVID19 dataset (Spain), which, while enhancing the dataset's size and regional representation, introduces a notable class imbalance. Specifically, COVID-19 positive cases (59,961) substantially outnumbered normal and non-COVID pneumonia cases (27,270). This disproportion primarily stems from the emphasis of public datasets on rapid COVID-specific data collection during the pandemic, which may skew model learning and diagnostic performance if not addressed.

Data Preprocessing

To ensure robust model performance across varying demographics, modalities, and clinical conditions, a comprehensive data preprocessing pipeline was applied. The steps undertaken effectively addressed initial issues of class imbalance, missing metadata, and image inconsistencies.

Class and Source Balancing

After applying undersampling and dropping countries with ultra-low samples, 3000 cases from Spain, 2949 from China, 1761 from the United States, 1526 from France, 1106 from Russia, and 718 from Iran were retained, with 7572 COVID cases and 3488 non-COVID cases.

Metadata Imputation Results

Age had 5537 missing values imputed using country-wise medians, and gender had 5511 missing values imputed using country-wise modes. Metadata completeness improved to 100%, allowing demographic-aware stratification and analysis during model evaluation.

Age and Gender Distribution

The age range was 0 to 100 years. After removing outliers, the age group distribution remained imbalanced, with adults (n=7219) forming the majority, followed by elderly (n=2234), young adults (n=1553), and children (n=54). The distribution reflects a population skew that may influence age-specific modeling outcomes, and stratified labels by age group ensure balanced data. Gender balance included 34.1% (3398/9954) males and 65.9% (6556/9954) females. After processing, the dataset had 4509 positive and 2047 negative cases among females, and 2307 positive and 1091 negative cases among males. This balanced demographic composition supports robust model evaluation across diverse patient profiles.

Dataset Overview After Balancing

After applying country-based filtering, undersampling, and augmentation, a more equitable distribution of samples across countries and classes was achieved. The total number of curated images was 11,052, with 8842 images in the training set and 2210 images in the validation set. Image dimensions were resized to 75×75 pixels with 3 RGB channels, and normalization was applied with all pixel values rescaled to the [0, 1] range.

Country-Level Balance (Postaugmentation)

Augmentation techniques were applied particularly to underrepresented classes to reduce class imbalance and enhance model generalization. A balanced representation (2034 COVID samples per country) was achieved across 6 key contributors (China, France, Iran, Russia, Spain, and the United States). Similarly, normal samples were balanced at 1249 images across the same regions, improving generalization across populations (Multimedia Appendix 4).

The dataset comprised 12,204 COVID-19–positive images and 7494 normal images, indicating a moderate class imbalance favoring positive cases. This distribution highlights the need for balancing techniques such as augmentation during model training.

Augmentation Impact

The applied augmentation techniques (flip, rotate, zoom, contrast, and translation) not only balanced the dataset but also increased image variability, simulating real-world noise and improving model resilience to unseen data (Multimedia Appendix 5). There was a nearly equal number of images per label (nearly 2000 per class) in each country, demonstrating successful class balancing to mitigate bias during model training.

Class Distribution (Postaugmentation)

There was an equal number of COVID-positive and normal (COVID-negative) images (12,204 each), reflecting the successful application of augmentation techniques to balance the dataset and prevent model bias due to class imbalance. Class distribution after augmentation is presented in Multimedia Appendix 6.

Modeling

To ensure a fair and consistent evaluation, all models were trained using standardized input settings. Each image was resized to 224×224 pixels, producing an input shape of (224, 224, 3) to accommodate RGB color channels. Although the images originated in RGB format, they were converted to grayscale during preprocessing and normalized to a range of [0, 1] for efficient convergence.

All transfer learning architectures were trained for 30 epochs, a setting chosen to balance computational efficiency with sufficient learning. A batch size of 128 was used to maintain stable updates across mini-batches. Additionally, a shuffle buffer size of 10,000 ensured randomness in the training data pipeline, reducing overfitting risks.

This consistent training configuration was applied across all models (VGG16, ConvNeXtTiny, ResNet50, EfficientNetB0, EfficientNetV2B0, DenseNet121, MobileNet, MobileNetV2, and NASNetMobile).

Through hyperparameter tuning, the DenseNet121 architecture was found to yield the best performance. Its final configuration included dropout layer 1 with 0.3, dense layer 1 with 128 units, a learning rate of 0.00037758, and a weight decay of 7.4855e-05. This architecture and training regime were optimized to prevent overfitting while maintaining high model generalization on unseen data.

Model Evaluation

Among the evaluated models, DenseNet121 delivered the best overall performance, achieving 98% accuracy, 96.8% precision, 98.8% recall, and an AUC of 0.998, indicating a well-balanced and highly effective binary classifier (Figure 1; Table 4). NASNetMobile and VGG16 also showed strong performance, with high scores across all metrics, making them solid alternatives. ResNet50 showed competitive results but fell slightly short of the top 3 models, particularly in precision. On the other hand, models, such as EfficientNetB0, EfficientNetV2B0, ConvNeXtTiny, and MobileNet, showed poor performance. Despite their perfect recall, their low precision and AUC values suggest that they overpredicted the positive class, leading to high false positive rates. MobileNetV2, despite a decent accuracy and AUC, failed to maintain balance across precision and recall, making it less suitable for reliable classification in this context. Given its superior and consistent results, DenseNet121 stands out as the most suitable model for deployment, offering both robustness and high predictive accuracy for this binary classification task.

Figure 1. The training and validation (A) accuracy and (B) loss curves of DenseNet121 (densely connected convolutional network-121 layers) over 30 epochs, showing strong learning convergence with minimal divergence between the training and validation sets, which is an indicator of effective generalization.

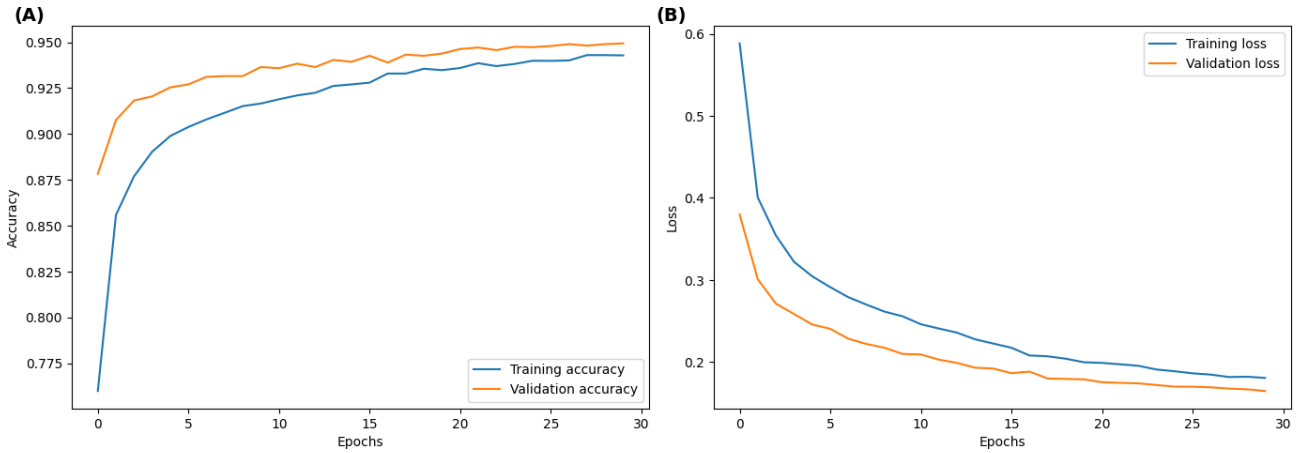


Table . Comparative analysis findings of performance metrics for all transfer learning models applied to the task of COVID-19 detection from medical images.

Model	Accuracy	Precision	Recall	F_1-score	AUC[a]
EfficientNet[b]B0	0.46219	0.46219	1.00000	0.63218	0.33122
EfficientNetV2B0	0.46219	0.46219	1.00000	0.63218	0.63435
MobileNet[c]	0.54306	0.50287	0.99545	0.66819	0.93267
ConvNeXtTiny[d]	0.46219	0.46219	1.00000	0.63218	0.50726
ResNet50[e]	0.92542	0.87885	0.97273	0.92341	0.99033
VGG16[f]	0.93487	0.91087	0.95227	0.93111	0.98431
NASNetMobile[g]	0.95798	0.93290	0.97954	0.95565	0.99619
MobileNetV2	0.97370	0.96874	0.97773	0.97321	0.97990
DenseNet121[h]	0.98004	0.96882	0.98864	0.97863	0.99830

[a]AUC: area under the receiver operating characteristic curve.

[b]EfficientNet: efficient network.

[c]MobileNet: mobile network.

[d]ConvNeXtTiny: convolutional next-tiny.

[e]ResNet50: residual network-50 layers.

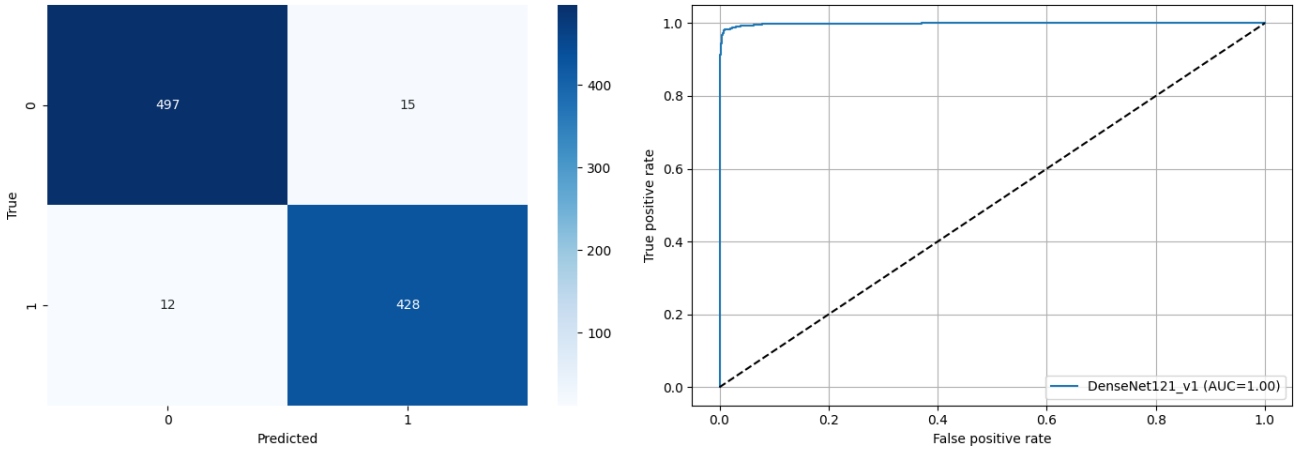
[f]VGG16: Visual Geometry Group network-16 layers.

[g]NASNetMobile: neural architecture search network-mobile version.

[h]DenseNet121: densely connected convolutional network-121 layers.

The confusion matrix reflects DenseNet121's exceptional classification accuracy with minimal misclassification (Figure 2A).

Figure 2. (A) Confusion matrix and (B) receiver operating characteristic curve for DenseNet121 (densely connected convolutional network-121 layers). AUC: area under the receiver operating characteristic curve.



The balance indicates that the model is not only highly accurate but also well-calibrated in terms of sensitivity (recall) and specificity.

The receiver operating characteristic curve further supports these results, with an AUC of 1.00, demonstrating near-perfect separation between positive and negative classes. The curve closely hugs the top-left corner, indicating an excellent tradeoff between the true positive rate and false positive rate (Figure 2B).

Together, these visualizations affirm DenseNet121's reliability and robustness for the binary classification task of COVID-19 detection, outperforming other evaluated architectures in both quantitative metrics and qualitative visual assessment.

Discussion

Summary

The results show that DenseNet121 achieved the highest performance, with 98% accuracy, 96.8% precision, and 98.8% recall, demonstrating robust diagnostic capabilities.

Conclusion

This study introduces a robust deep learning framework for COVID-19 diagnosis using chest X-ray and CT imaging, emphasizing both high model performance and real-world deployment feasibility. Leveraging imaging data from 19 countries across diverse age groups, genders, and COVID-19 variants, the study used comprehensive preprocessing, undersampling, and data augmentation techniques to ensure balanced and representative datasets. To ensure practical deployment, models were optimized through quantization and pruning, making them lightweight and suitable for web-based diagnostic platforms via cloud APIs (Flask or RESTAPI with TensorFlow Serving) and mobile apps using TensorFlow Lite or ONNX for on-device diagnosis, which can be especially valuable in low-resource and rural settings. The framework further integrates Grad-CAM visualizations for explainability, federated learning for privacy-preserving collaboration across hospitals, and longitudinal monitoring for tracking long COVID or reinfection cases. These features collectively position the system as a clinically relevant, mutation-resilient, and scalable solution for COVID-19 screening and triage in modern health care environments. For future work, there is an aim to extend this framework to multiclass classification, distinguishing between lung pathologies such as tuberculosis, AIDS, and COVID-19. This initiative will be pursued in collaboration with clinicians to enhance diagnostic specificity and clinical utility.

Future Work

Clinical Validation Across Institutions

There is an aim to collaborate with multiple hospitals and diagnostic centers to externally validate the model on institution-specific datasets. This will help assess the model's generalizability and robustness across different scanners, protocols, and patient populations.

Integration With EHRs

Work is underway to integrate the diagnostic tool with EHR systems for seamless access to patient history and real-time imaging data, enabling context-aware predictions and decision support.

Deployment on Web and Mobile Platforms

The final model is being optimized using techniques, such as quantization and pruning, for deployment on edge devices and cloud platforms. This will support real-time diagnosis via a web interface and mobile app, particularly in resource-constrained or rural areas.

Regulatory Readiness and Clinical Trials

Documentation and performance benchmarks are being prepared to pursue regulatory approval (Conformité Européenne marking and Food and Drug Administration clearance). A prospective clinical trial is also being designed to measure diagnostic impact in a real-world setting.

Extension to Long COVID and Follow-Up Monitoring

There is a plan to adapt the system for longitudinal analysis, enabling clinicians to track radiological changes over time, which can be useful for monitoring long COVID progression or reinfections.

Federated Learning for Privacy-Preserving AI

To support data privacy and multi-institutional collaboration, an attempt will be made to explore federated learning frameworks that allow model training on decentralized data without sharing patient images.

Acknowledgments

I would like to express my gratitude to my supervisor Li Zhang who shaped, guided, and refined my work through this experiment. Her subject expertise, intuition on the areas to explore, and patience as a teacher played a major part in making this project what it is today.

Data Availability

The full implementation and the pretrained models are publicly available on GitHub [32].

Conflicts of Interest

None declared.

Multimedia Appendix 1
Distribution of COVID-positive and normal chest images by country.
[PNG File, 24 KB - xmed_v6i1e75015_app1.png]

Multimedia Appendix 2
Source-wise distribution of imaging data used in the study.
[PNG File, 19 KB - xmed_v6i1e75015_app2.png]

Multimedia Appendix 3
Label-wise distribution of COVID-positive and negative cases across various data sources.

[PNG File, 23 KB - xmed_v6i1e75015_app3.png]

Multimedia Appendix 4
Bar chart of image count per label and country after data augmentation, illustrating a balanced distribution of COVID and normal images across 6 countries, which ensured class uniformity for training deep learning models.
[PNG File, 27 KB - xmed_v6i1e75015_app4.png]

Multimedia Appendix 5
Bar chart of image count per label and country, showing the distribution of COVID-19 and normal images across 6 countries after data augmentation.
[PNG File, 39 KB - xmed_v6i1e75015_app5.png]

Multimedia Appendix 6
Bar chart of the total image count per label after augmentation.
[PNG File, 17 KB - xmed_v6i1e75015_app6.png]

References

1. Pneumonia of unknown cause – China. World Health Organization. 2020. URL: https://www.who.int/emergencies/disease-outbreak-news/item/2020-DON229 [accessed 2025-09-12]
2. Coronavirus disease (COVID-19) - Overview. World Health Organization. URL: https://www.who.int/health-topics/coronavirus#tab=tab_1 [accessed 2025-09-12]
3. COVID-19: new variants in 2025. Ada Health. 2025. URL: https://ada.com/covid/what-strain-of-covid-is-going-around/ [accessed 2025-09-12]
4. Weekly epidemiological update on COVID-19 - 25 August 2023. World Health Organization. 2023. URL: https://www.who.int/publications/m/item/weekly-epidemiological-update-on-covid-19---30-august-2023 [accessed 2025-09-12]
5. COVID-19 Public Health Emergency of International Concern (PHEIC) Global research and innovation forum. World Health Organization. 2020. URL: https://www.who.int/publications/m/item/covid-19-public-health-emergency-of-international-concern-%28pheic%29-global-research-and-innovation-forum [accessed 2025-09-12]
6. COVID-19 - global situation. World Health Organization. 2025. URL: https://www.who.int/emergencies/disease-outbreak-news/item/2025-DON572 [accessed 2025-09-12]
7. Katella K. 3 things to know about XEC, the dominant COVID strain. Yale Medicine. 2024. URL: https://www.yalemedicine.org/news/3-things-to-know-about-xec-the-latest-covid-strain [accessed 2025-09-12]
8. Surveillance and data analytics. Centers for Disease Control and Prevention. 2025. URL: https://covid.cdc.gov/covid-data-tracker [accessed 2025-09-12]
9. WHO's 2025 updates on COVID-19 variants: focus on XEC, testing, and recovery. ASSURE. 2025. URL: https://assure-test.com/2025/02/05/whos-2025-updates-on-covid-19-variants-focus-on-xec-testing-and-recovery/ [accessed 2025-09-12]
10. Tracking SARS-CoV-2 variants. World Health Organization. 2025. URL: https://www.who.int/activities/tracking-sars-cov-2-variants [accessed 2025-09-12]
11. Coronavirus disease (COVID-19) - Symptoms. World Health Organization. URL: https://www.who.int/health-topics/coronavirus#tab=tab_3 [accessed 2025-09-12]
12. COVID-19 symptoms: Omicron vs. Delta. Ada Health. 2025. URL: https://ada.com/covid/covid-19-omicron-vs-delta-symptoms/ [accessed 2025-09-12]
13. Ullah SMA, Islam MM, Mahmud S, Nooruddin S, Raju S, Haque MR. Scalable telehealth services to combat novel coronavirus (COVID-19) pandemic. SN Comput Sci 2021;2(1):18. [doi: 10.1007/s42979-020-00401-x] [Medline: 33426530]
14. Wang L, Lin ZQ, Wong A. COVID-Net: a tailored deep convolutional neural network design for detection of COVID-19 cases from chest x-ray images. Sci Rep 2020 Nov 11;10(1):19549. [doi: 10.1038/s41598-020-76550-z] [Medline: 33177550]
15. Reshan MSA, Gill KS, Anand V, et al. Detection of pneumonia from chest x-ray images utilizing MobileNet model. Healthcare (Basel) 2023 May 26;11(11):1561. [doi: 10.3390/healthcare11111561] [Medline: 37297701]
16. Mujahid M, Rustam F, Álvarez R, Luis Vidal Mazón J, Díez IDLT, Ashraf I. Pneumonia classification from x-ray images with Inception-V3 and Convolutional Neural Network. Diagnostics (Basel) 2022 May 21;12(5):1280. [doi: 10.3390/diagnostics12051280] [Medline: 35626436]
17. Ghaderzadeh M, Asadi F, Jafari R, Bashash D, Abolghasemi H, Aria M. Deep convolutional neural network-based computer-aided detection system for COVID-19 using multiple lung scans: design and implementation study. J Med Internet Res 2021 Apr 26;23(4):e27468. [doi: 10.2196/27468] [Medline: 33848973]
18. Jiang W, Ji W, Zhang Y, et al. An update on detection technologies for SARS-CoV-2 variants of concern. Viruses 2022 Oct 22;14(11):2324. [doi: 10.3390/v14112324] [Medline: 36366421]

19. Miró Catalina Q, Fuster-Casanovas A, Solé-Casals J, Vidal-Alaball J. Developing an artificial intelligence model for reading chest x-rays: protocol for a prospective validation study. JMIR Res Protoc 2022 Nov 16;11(11):e39536. [doi: 10.2196/39536] [Medline: 36383419]
20. Wang Q, Mellis IA, Ho J, et al. Recurrent SARS-CoV-2 spike mutations confer growth advantages to select JN.1 sublineages. Emerg Microbes Infect 2024 Dec;13(1):2402880. [doi: 10.1080/22221751.2024.2402880] [Medline: 39259045]
21. Scott Mader K. The Lung Image Database Consortium image collection (LIDC-IDRI). IEEE DataPort. 2021. URL: https://ieee-dataport.org/documents/lung-image-database-consortium-image-collection-lidc-idri [accessed 2025-09-12]
22. SIRM - Società Italiana di Radiologia Medica e Interventistica. URL: https://sirm.org [accessed 2025-09-12]
23. BIMCV-COVID19, Conjuntos de datos relacionados con el curso de patología de COVID19 [Article in Spanish]. BIMCV. 2023. URL: https://bimcv.cipf.es/bimcv-projects/bimcv-covid19/ [accessed 2025-09-12]
24. iCTCF: CT images and clinical features for COVID-19. National Genomics Data Center - China National Center for Bioinformation. URL: https://ngdc.cncb.ac.cn/ictcf/ [accessed 2025-09-12]
25. CT images in COVID-19. The Cancer Imaging Archive. URL: https://www.cancerimagingarchive.net/collection/ct-images-in-covid-19/ [accessed 2025-09-12]
26. MIDRC-RICORD-1A. The Cancer Imaging Archive. URL: https://www.cancerimagingarchive.net/collection/midrc-ricord-1a/ [accessed 2025-09-12]
27. MIDRC-RICORD-1B. The Cancer Imaging Archive. URL: https://www.cancerimagingarchive.net/collection/midrc-ricord-1b/ [accessed 2025-09-12]
28. MIDRC-RICORD-1C. The Cancer Imaging Archive. URL: https://www.cancerimagingarchive.net/collection/midrc-ricord-1c/ [accessed 2025-09-12]
29. STOIC2021 - COVID-19 AI Challenge. STOIC2021 Grand Challenge. URL: https://stoic2021.grand-challenge.org/ [accessed 2025-09-12]
30. COVID-19. Radiopaedia. URL: https://radiopaedia.org/articles/covid-19-4 [accessed 2025-09-12]
31. Datasets [Article in Russian]. Center of Diagnostics and Telemedicine. URL: https://mosmed.ai/en/datasets/ [accessed 2025-09-12]
32. Dharmik A. COVID-19-APP. GitHub. 2025. URL: https://github.com/AnjaliDharmik/COVID-19-APP [accessed 2025-09-12]

Abbreviations

AI: artificial intelligence
API: application programming interface
AUC: area under the receiver operating characteristic curve
BIMCV-COVID19: Banco de Imágenes Médicas de la Comunidad Valenciana–COVID-19
CNCB: China National Center for Bioinformation
CNN: convolutional neural network
ConvNeXtTiny: convolutional next-tiny
CT: computed tomography
DenseNet121: densely connected convolutional network-121 layers
EfficientNet: efficient network
EHR: electronic health record
iCTCF: CT images and clinical features for COVID-19
LFD: lateral flow device
LIDC-IDRI: Lung Image Database Consortium image collection
MIDRC-RICORD: Medical Imaging Data Resource Center - RSNA International COVID-19 Open Radiology Database
MobileNet: mobile network
NASNetMobile: neural architecture search network-mobile version
RAT: rapid antigen test
RestNet50: residual network-50 layers
RT-PCR: reverse transcription–quantitative polymerase chain reaction
SIRM: Società Italiana di Radiologia Medica e Interventistica
STOIC: Study of Thoracic CT in COVID-19
TCIA: The Cancer Imaging Archive
VGG16: Visual Geometry Group network-16 layers
WHO: World Health Organization

Edited by F Wu; submitted 26.03.25; peer-reviewed by E Ndezure, I Odezuligbo, CLA Sunny; revised version received 16.08.25; accepted 29.08.25; published 26.09.25.

Please cite as:
Dharmik A
COVID-19 Pneumonia Diagnosis Using Medical Images: Deep Learning–Based Transfer Learning Approach
JMIRx Med 2025;6:e75015
URL: https://xmed.jmir.org/2025/1/e75015
doi:10.2196/75015

Real-World Performance of COVID-19 Antigen Tests: Predictive Modeling and Laboratory-Based Validation

Miguel Bosch[1,2], PhD; Dawlyn Garcia[2], MSc; Lindsey Rudtner[2], BSc; Nol Salcedo[2], MSc; Raul Colmenares[1], MSc; Sina Hoche[2], PhD; Jose Arocha[2], MSc; Daniella Hall[2], PhD; Adriana Moreno[1], MSc; Irene Bosch[2], PhD

[1]Info Analytics Innovations, Houston, TX, United States

[2]IDX20 Inc, 166 Clinton Rd, Brookline, MA, United States

Corresponding Author:
Irene Bosch, PhD
IDX20 Inc, 166 Clinton Rd, Brookline, MA, United States

Related Articles:

Companion article: https://www.medrxiv.org/content/10.1101/2024.10.21.24315762v1

Companion article: https://med.jmirx.org/2025/1/e83476

Companion article: https://med.jmirx.org/2025/1/e83479

Companion article: https://med.jmirx.org/2025/1/e83474

Abstract

Background: Rapid and safe deployment of lateral-flow antigen tests, coupled with uncompromised quality assurance, is critical for outbreak control and pandemic preparedness, yet real-world performance assessment still lacks laboratory and quantitative approaches that remain uncommon in current regulatory science. The approach proposed here can help standardize and accelerate early phase appraisal of antigen tests in preparation for clinical validation.

Objective: The aim of this study is to present a quantitative, laboratory-anchored framework that links image-based test line intensities and the population distribution of naked-eye limits of detection (LoD) to a probabilistic prediction of positive percent agreement (PPA) as a function of viral-load–related variables (eg, quantitative real-time polymerase chain reaction [qRT-PCR] cycle thresholds [Cts]). Using dilution-series calibrations and a Bayesian model, the predicted PPA-vs-Ct curve closely tracks the observed PPA in a real-world self-testing cohort.

Methods: The proposed methodology combines: (1) a quantitative evaluation of the test signal response to concentrations of target protein and inactive virus or active virus, (2) a statistical characterization of the LoD using the observer's visual acuity of the test band, and (3) a calibration of a gold-standard method (eg, qRT-PCR cycles) against virus concentration. We elaborate these quantitative methods and unfold a Bayesian-based predictive model to describe the real-world performance of the antigen test, quantified by the probability of positive agreement as a function of viral-load variables like qRT-PCR Cts.

Results: We applied the methodology by characterizing each brand of COVID-19 antigen test and estimating its real-world probability of agreement with qRT-PCR. We aligned protein and inactivated-virus standard curves at matched signal intensities and fit a linear calibration linking protein to viral concentrations. Using logistic regression, we modeled the PPA as a continuous function of qRT-PCR Ct, then integrated this curve over a predefined reference Ct distribution to obtain the expected sensitivity. This standardization enables consistent performance comparisons across sites.

Conclusions: Modeling performance under real-world conditions requires coupling laboratory evaluation with the population's ability to perceive the test's visual signal. We represent observer capability as a probability density function of the LoD over the signal-intensity domain. Rather than reporting bin-based sensitivity, we summarize performance with the PPA as a continuous function of qRT-PCR Ct. Our framework produces PPA–Ct curves by composing (1) normalized signal-to-concentration models from the laboratory, (2) the observer LoD distribution, and (3) a Ct-to-viral-load calibration. The resulting inferences are inherently context-bound—disease-, assay-, and setup-specific. External validity depends on the particular antigen lateral-flow test, the user population (visual acuity and interpretation), and cross-laboratory qRT-PCR calibration. Comprehensive clinical studies under intended-use conditions are still required before making generalized claims.

Trial Registration: ClinicalTrials.gov NCT05884515; https://clinicaltrials.gov/study/NCT05884515

(JMIRx Med 2025;6:e68376) doi:10.2196/68376

KEYWORDS

COVID-19; SARS-CoV-2 antigen test; lateral flow assay; point-of-care diagnostics; real-world performance; Langmuir–Freundlich isotherm; Bayesian regression (Monte Carlo); probability of positive agreement; limit of detection; image-based signal quantification

Introduction

Quantifying the performance of antigen lateral flow tests (Ag-LFT) according to the US regulatory science standards commonly requires the calculation of test performance statistics—for example, sensitivity or positive percent agreement (PPA) of an antigen test's (AT) binary assessments with reference to the quantitative real-time polymerase chain reaction (qRT-PCR) gold standard [1-3]. These statistics are based on human clinical samples, requiring paired qRT-PCR cycle thresholds (Cts) to generate performance data and corresponding test validations as described in a comprehensive literature review for COVID-19 studies [4]. The clinical performance of Ag-LFTs increases at higher viral loads (low Ct values), which present early on in the symptom window, and test performance declines with low viral loads (high Ct values) at the end of the acute disease window [2,5-10]. These realities motivate a fast, laboratory-anchored, model-based appraisal that anticipates PPA-vs-Ct before large trials. That is, AT performance is closely related to the sample viral load and the performance statistics are dependent on several factors such as viral load distribution, specific virus variants [8,9,11], and symptomatic versus asymptomatic cases [7,12], as well as the observer's training [6,13-17]. Hence, the regulatory process is typically a lengthy process and includes appropriate sample size and requires a viral load distribution to cover the spectrum of target concentrations.

Because Ag-LFTs are powerful tools for transmission control and epidemic mitigation, the fast and safe deployment of tests without compromising quality assurance in the evaluation process is key for outbreak control and pandemic preparedness. Home testing and easy access to Ag-LFTs enables them to be used for serial testing, as has been recommended. For COVID-19 Ag-LFTs, serial testing increases effective sensitivity [18], including in home testing [7,13,19].

We have developed a methodology for the quantitative evaluation of SARS-CoV-2 ATs based on laboratory measurements of the regions of interest (ROIs), including the test's regions and corresponding normalized signal intensity and binary naked-eye user assessments for positive or negative results. In both cases, we characterized the test performance according to the sample concentration of the target recombinant protein, with heat-inactivated virus as well as biologically active virus, and we used human samples self-collected 2 times per week, under a prospective clinical protocol (ClinicalTrials.gov: NCT05884515). To support accurate self-reporting, participants received a brief, standardized orientation with visual aids on test interpretation, photo capture, and upload procedures. Including the statistical characterization of the user population's limit of detection (LoD) in the signal intensity domain, we developed a predictive model for the probability of positive agreement in real-world conditions.

Our method involves (1) characterizing the AT signal intensity with protein and inactivated virus dilutions, (2) calibrating the qRT-PCR cycles with virus dilutions, (3) characterizing the signal intensity LoD of the user population for the AT, and (4) predicting the real-world probability of a positive agreement signal response of the AT. Our methods have the advantage of being formulated using continuous variable analysis and probability models instead of plain discrete analysis and sample statistics.

We demonstrate our methodology capabilities when comparing the predicted probability of positive agreement with that generated using real-world data collected through an institutional review board (IRB)–approved study for frequent antigen testing to monitor COVID-19 in an underserved population (ClinicalTrials.gov: NCT05884515). Participants consisted of individuals from vulnerable populations in low-income and assisted-living facilities located in the city of Chelsea, MA. Recruitment included people living in state-regulated, independent senior living communities and other residents of Chelsea. The consented participants were provided with ATs to routinely self-test for COVID-19 at home or in community centers, 2 times per week, uploading the test results and photos to the project informatics platform. We obtained confirmatory qRT-PCR data for all positive results detected by the home AT and for a random number of negative results from an independent Clinical Laboratory Improvement Amendments laboratory. The certified Clinical Laboratory Improvement Amendments laboratory procedures were approved by the IRB and the qRT-PCR data shared included the Ct values for each submitted test.

We describe the quantitative analysis of the AT for signal intensity and naked-eye binary data, the characterization of the user's LoD, the calibration of the qRT-PCR, and the formulation of the predictive model. In the Results section, we illustrate the application of the described methodology with the characterization of an AT in the common cassette device presentation and compare the predicted result with the real-world probability of positive agreement.

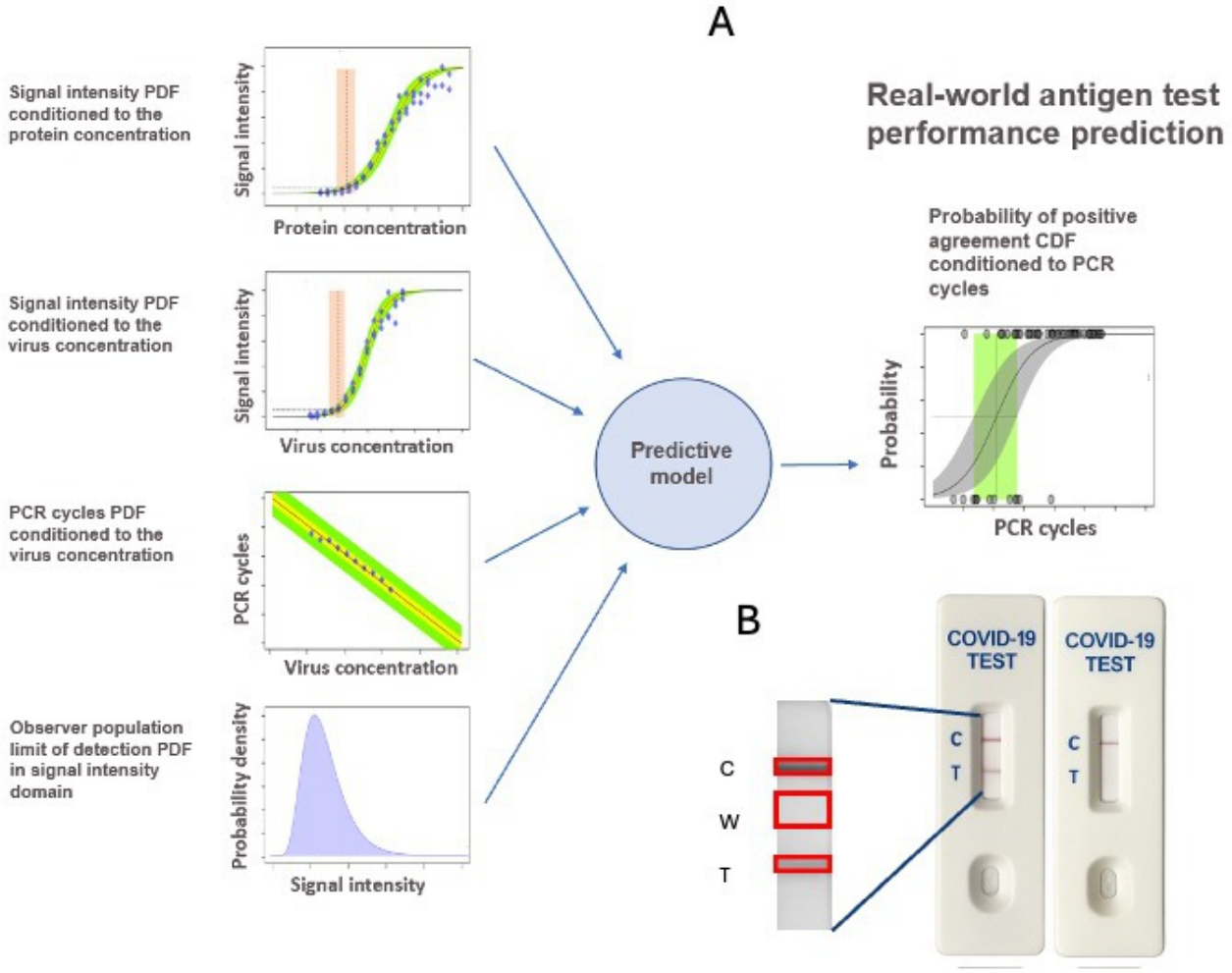
Methods

Overview

The present methodology combines (1) a quantitative evaluation of the test signal response to concentrations of target protein and inactive or active virus, (2) a statistical characterization of the LoD using observers' visual acuity of the test band, and (3) a calibration of a gold-standard method (ie, qRT-PCR cycles) against virus concentration. We elaborate these quantitative methods and unfold a Bayesian-based predictive model to describe the real-world performance of the AT, quantified by the probability of positive agreement as a function of viral-load

variables like qRT-PCR Cts. Figure 1A describes the different types of information involved in the predictive model.

Figure 1. (A) Schematic of the inference model. Input components are based on various laboratory calibrations of the antigen test–normalized signal intensity against protein and/or virus concentrations, qRT-PCR Cts against virus concentrations, and the limit of detection PDF of the observer population's visual assessment of the test result. The model predicts performance for real-world conditions of the antigen tests, quantified as the probability of agreement function. (B) Schematic of the antigen test results as positive or negative (recombinant antigen and/or inactivated virus or patient sample), according to the manufacturer's protocol and read at the specified development time. The cassette image was generated using a cell phone camera, and the annotated regions of interest were the control band (C), local white background (W), and test band (T) areas. Mean gray scale values from these regions of interest are used to compute normalized signal intensity for model fitting. Ct: cycle threshold; PDF: probability density function; PPA: positive percent agreement; qRT-PCR: quantitative real-time polymerase chain reaction.



AT Signal Response Characterization

In a selected lateral flow test, we measured the signal intensity of the test, white background, and control band, captured digitally in a photograph using a cell phone camera (Figure 1B). We processed this image and evaluated the average pixel intensity in 3 ROIs on the nitrocellulose strip. The resulting gray scale–normalized signal intensity is a continuous variable independent of the observer. There are existing methodologies to assess AT performance, such as test band signal intensity [20] and LoD studies [21]. We calculated the signal intensity by subtracting the white background and test band average pixel brightness and normalizing by the largest signal intensity present in the dataset. The software was designed to provide use instructions, obtain gray scale pixel intensity of the ROI, compute normalized pixel intensities, and generate a written report [22].

We analyzed the signal intensity response of an AT, calculating the signal intensity across a dilution series of the target recombinant protein. To fit these data, we used techniques based on isotherm modeling [23,24] and used the Langmuir-Freundlich adsorption model [25,26],

$$I = k\,C^{b} / (1 + k\,C^{b}) \quad (1)$$

with I being the normalized signal intensity, C the concentration, k the adsorption equilibrium constant, and b an empirical exponent close to one. The model parameters to estimate by fitting the normalized intensity data corresponded to the adsorption constant k and the exponent b. We used a Bayesian regression solved with Monte Carlo sampling, which provided a description of the model uncertainties. We followed a similar procedure to characterize the relationship of the normalized signal intensity with other variables in addition to that of using recombinant protein concentrations (eg, with a known plaque-forming unit/mL of SARS-CoV-2 followed by chemical inactivation of the virus).

Probability of Agreement Function in Naked-Eye Assessment

A common use of ATs involves human naked-eye interpretation of the result. The outcome of each assessment is a binary variable, either positive or negative (1 for positive or 0 for negative for mathematical analysis). We considered that for the naked-eye analysis, individuals required training to follow specific protocols to properly report test results. Before self-testing, participants completed a brief, standardized orientation (10 - 15 min) delivered in English or Spanish according to specifications of the Chelsea study. The module covered (1) correct nasal specimen collection and adherence to the manufacturer's instructions for use, (2) strict timing of the development/read window, (3) interpretation of positive, negative, and invalid outcomes, and (4) photographing and uploading results from a cell phone camera or tablet to the study's digital platform in real time. Participants received a 1-page pictorial quick-start guide. In addition, refresher prompts were available on the study platform. Thereby, efforts were made to stabilize the observer LoD distribution data used in our model.

The PPA or sensitivity of a test is a well-known measure of test performance. The PPA is strongly dependent on the viral concentration distribution of the tested samples (eg, sensitivity improves with higher concentrations of the target).

For an accurate description of the naked-eye performance of the test, we estimated the PPA as a *function* of the nucleoprotein concentration or other viral concentration–related variable, such as qRT-PCR Cts [8]. We modeled the PPA with a logistic function,

$$p(x)=\frac{1}{1+e^{-(a+bx)}} \quad (2)$$

with x being the viral-load–related variable and p(x) being the probability of positive agreement function. The model parameters to estimate fitting the binary naked-eye data are the intercept a and the slope b. The PPA function is commonly described against qRT-PCR cycles; similarly, we applied this method for other viral-load–related variables (like concentration and normalized signal intensity).

Predictive Model for the Probability of Positive Agreement

The formulation of the model followed a probabilistic approach, meaning that variables and relationships across the model were randomized and described by probability density functions (PDFs). We defined random variables used in this formulation. There was a group of continuous positive variables that were related to the viral load: the recombinant protein concentration xprot, the virus concentration xvir, the test–normalized signal intensity xint, the observer LoD–normalized signal intensity xlod, and the qRT-PCR cycles xcycle. In addition, we had the binary agreement variable *A*, which indicated the observer assessment of the test outcome, with values of 0 (for negative) or 1 (for positive).

For the purpose of this analysis, the LoD did not represent an exact value. We analyzed the LoD associated with a group of observers (ie, a certain population) or a single observer; LoD depends on different environmental circumstances (eg, illumination, visual context) and individual abilities. Hence, we consider the LoD as a random variable defined by a PDF pLoD(xlod) in the domain of the normalized signal intensity xint. The probability of positive agreement (ie, the conditional positive agreement PDF) was the corresponding cumulative distribution function of the LoD PDF,

$$p(A=1|\,xint)=\int_0^{xint} pLoD(xlod)\;dxlod \quad (3)$$

Correspondingly, the LoD PDF in the signal intensity domain was the derivative of the probability of positive agreement in the same domain. It summarized the process of observation and assessment of the testing device by the observer or the observer population.

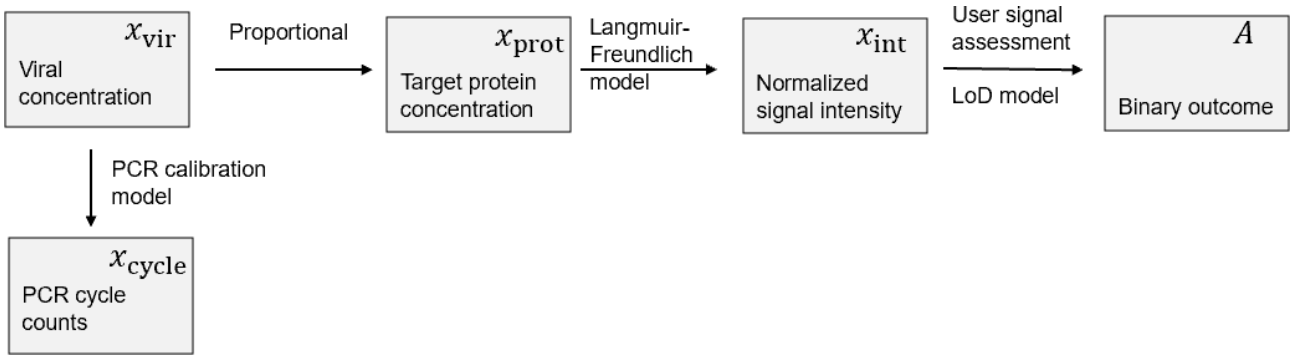
Probability of Positive Agreement Across Viral-Load–Related Domains

To follow, we transformed the probability of agreement in the signal intensity domain (expression 3) to the rest of the viral-load–related domains: recombinant protein concentration, viral concentration, and qRT-PCR cycles. The defined random variables and their causal dependencies are described in the Bayesian network of Figure 2. Let us consider, first, the case of the protein concentration domain. As per the intensity analysis described previously, we experimentally estimated a conditional probability for the signal intensity given the recombinant protein concentration pxint|xprot. We used this information to propagate the probability of positive agreement from the signal intensity domain to the protein concentration domain. For this purpose, we applied the probability chain rule,

$$p(A, xint \mid xprot)=p(A \mid xint, xprot)\;p(xint \mid xprot) \quad (4)$$

Integrating in the xint domain and taking into account that the observer assessment is only dependent on normalized signal intensity (see Figure 2) p(A | xint , xprot)=p(A | xint),

Figure 2. Bayesian network showing (boxes) the model random variables, (arrows) their causal relations and (annotations) relation models. LoD: limit of detection; PCR: polymerase chain reaction.



. (5)p(A ∣ xprot)=∫p(A ∣ xint) p(xint ∣ xprot) dxint

The functions within the integral involve the PPA in the signal intensity domain (expression 3) and the PDF of signal intensity conditioned to the protein concentration. We determined the integral by Monte Carlo integration. Similarly, we transformed the probability of agreement to the virus concentration domain and qRT-PCR cycles domain as

(6)p(A ∣ xvir)=∫p(A ∣ xint)p(xint ∣ xprot)p(xprot ∣ xvir)dxintdxprot

and

.

(7)p(A ∣ xcycle)=∫p(A ∣ xint)p(xint ∣ xprot)p(xprot ∣ xvir)p(xvir ∣ xcycle)dxintdxprotdxvir

To model the probability of agreement in the domain of qRT-PCR cycles, as is common for real-world testing, we solved Equation 7 by Monte Carlo integration. For this purpose, we needed to estimate models for the 4 conditional probabilities within the right-hand integrand, which involve the information represented in Figures 1 and 2. In our implementation, we first integrated in the virus concentration domain to have a relationship between the protein concentration and the qRT-PCR cycles, p(xprot | xcycle). Thus,

.

(8)p(A ∣ xcycle)=∫p(A ∣ xint)p(xint ∣ xprot)p(xprot ∣ xcycle)dxintdxprot

The conditional probability of positive agreement pA|xcycle fully describes the test performance in the qRT-PCR cycle domain. For a given PDF of the sample polymerase chain reaction (PCR) cycles distribution, the resulting sensitivity pA is by integration,

. (9)p(A=1)=∫p(A=1 ∣ xcycle) p(xcycle) dxcycle

For a given collection of N real-world samples with PCR cycles xcycle={x1 , x2 ,…xn ,…, xN } , the previous integral is approximated by the average of the PPA function evaluated at the sample qRT-PCR cycles,

. (10)p(A=1) ≈ 1N ∑1Np(A=1 | xn)

We have made software available to calculate the probability of agreement [22].

Ethical Considerations

This study was reviewed and approved by Advarra IRB for the protocol “Center of Complex Interventions – IDx20-001, Community frequent antigen testing to monitor COVID-19 in senior public housing setup (Pro00059157).” The most recent continuing review approval was granted on November 13, 2023, with an approval period through November 13, 2024. Advarra attests compliance with the US Department of Health and Human Services 45 CFR 46 and Food and Drug Administration 21 CFR 50/56 and is registered with OHRP/FDA (IRB number 00000971). All procedures adhered to the ethical standards of the responsible institutional/national committees and the World Medical Association Declaration of Helsinki. All personnel received certification and training through the Collaborative Institutional Training Initiative for human subjects research protection. There was no compensation to participants.

All study personnel completed Collaborative Institutional Training Initiative coursework prior to engaging in any human-subjects activities. This training is required by our IRB and institutional policy and included role-appropriate modules in Biomedical Human Subjects Research, Good Clinical Practice, Responsible Conduct of Research, Conflicts of Interest, and Health Insurance Portability and Accountability Act Privacy/Security. Certificates were verified by the private investigator and maintained on file. The curriculum covers the Belmont Report principles and applicable regulations (45 CFR 46 and Food and Drug Administration 21 CFR Parts 50/56), informed consent and documentation, recruitment and equitable selection, protection of vulnerable populations, adverse event and deviation reporting, data privacy/confidentiality, and secure data management.

Because the protocol involves point-of-care antigen testing and handling of respiratory specimens, staff also completed biosafety/Blood borne Pathogen training and followed BSL-2-appropriate standard operating procedures. All participant-facing procedures (screening, consent, anterior-nares swab collection, test execution, results disclosure) were conducted only by trained personnel under IRB-approved standard operating procedures. Study data were coded with limited identifiers, stored in access-controlled databases, and managed according to least-privilege access and audit-trail requirements. This statement documents personnel competence and compliance with human-subjects protections for the conduct of this study.

Prior to study procedures, all participants were informed of the study purpose, procedures, potential risks and benefits, data uses, and their right to withdraw without penalty. Written informed consent was obtained from each participant using IRB-approved consent materials. No identifiable personal information is reported in this manuscript.

Data were collected and stored using IRB-approved procedures designed to protect participant privacy and confidentiality; only deidentified or aggregated data are presented. For studies of internet/digital tools, we complied with applicable local, national, and international regulations on the protection of personal information, privacy, and human rights.

Any protocol amendments, consent-form changes, or substantive reportable events (eg, unanticipated problems, adverse device effects, or protocol violations affecting rights, safety, or data integrity) were submitted to Advarra in accordance with IRB requirements prior to implementation.

Results

We illustrate the application of the described methodology to characterize the analyzed COVID-19 AT brand and predict the corresponding real-world probability of agreement against qRT-PCR data.

Figure 3 shows the signal intensity data corresponding to protein dilutions prepared in the laboratory for this AT and the corresponding Langmuir-Freundlich regression model; from the analysis, we modeled the conditional PDF pxint|xprot. Figure 4 illustrates our modeled relationships across various viral-load–related variables based on our experimental characterization of the AT and qRT-PCR calibration. Figure 4A shows the signal intensity analysis of the AT based on serial dilutions of inactivated virus. The plot is similar to Figure 3, which shows the signal response to protein dilutions. By combining the protein and virus curves for common signal intensity responses, we calibrated a linear model that describes the relationship between protein and inactivated virus concentration (Figure 4B). Figure 4C shows the calibration of the qRT-PCR Ct curve based on PCR-analyzed inactivated virus dilution series. The qRT-PCR analysis of the dilution series was conducted by the same center used for self-testing our ordinary (real-world) testing program. Figure 4D shows the relationship between qRT-PCR cycles and protein concentration, transformed from the viral concentration domain by the protein-virus relationship characterized in Figure 4B. All the relationships shown in Figure 4 are modeled as conditional PDFs; for illustration, the plots show specific confidence limits.

Figure 3. Normalized signal intensity data for 3 series of protein dilution curves for one of the COVID-19 test brands and Langmuir-Freundlich model. The estimation of the LoD (LoD confidence intervals at 95%). LoD in signal intensity was 5% in the normalized signal intensity. LoD: limit of detection.

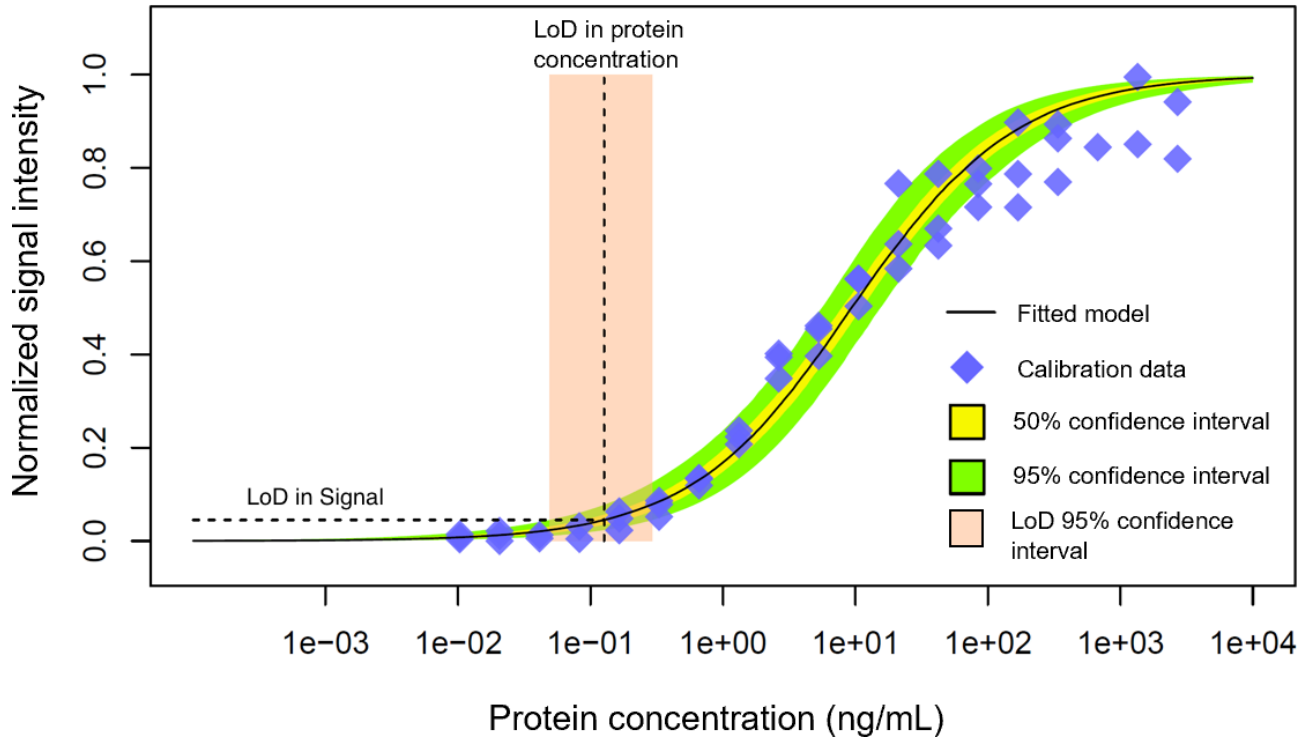
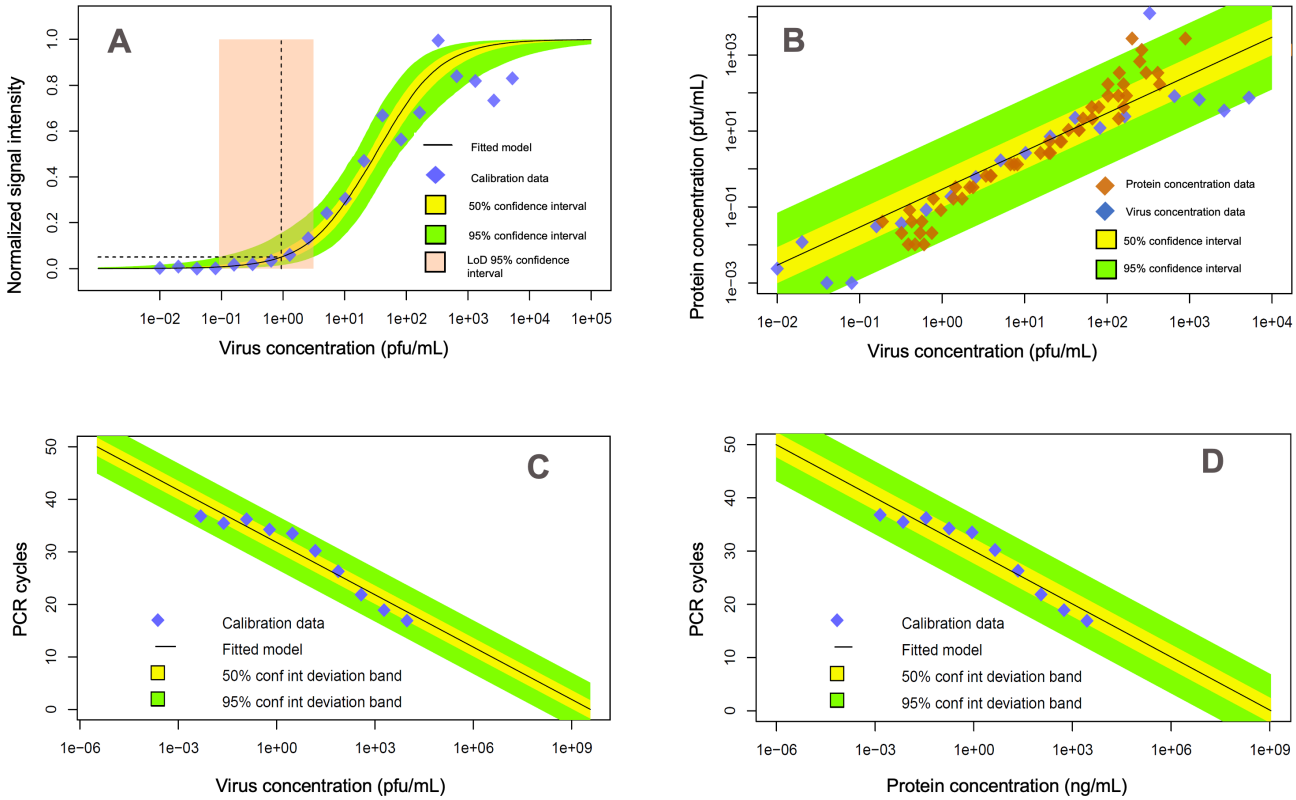
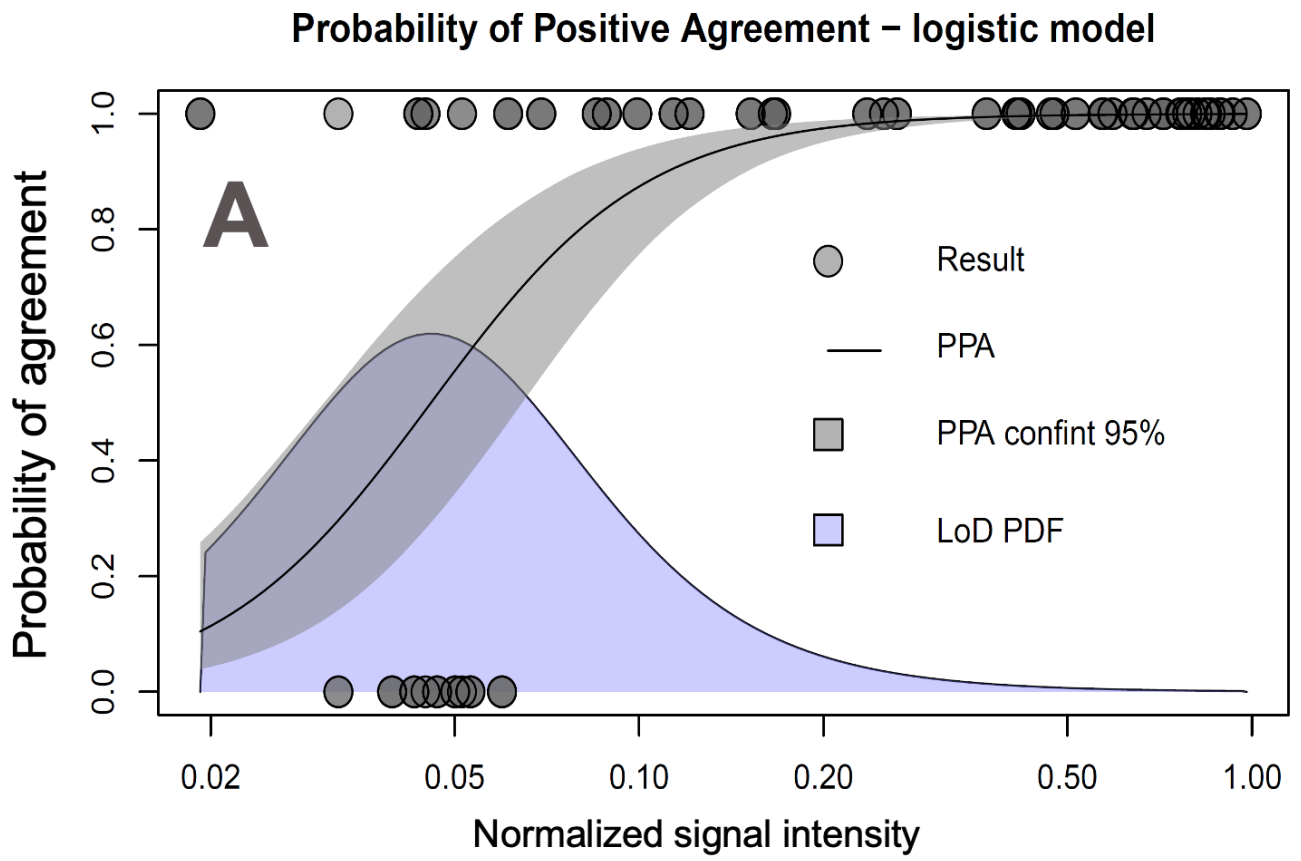


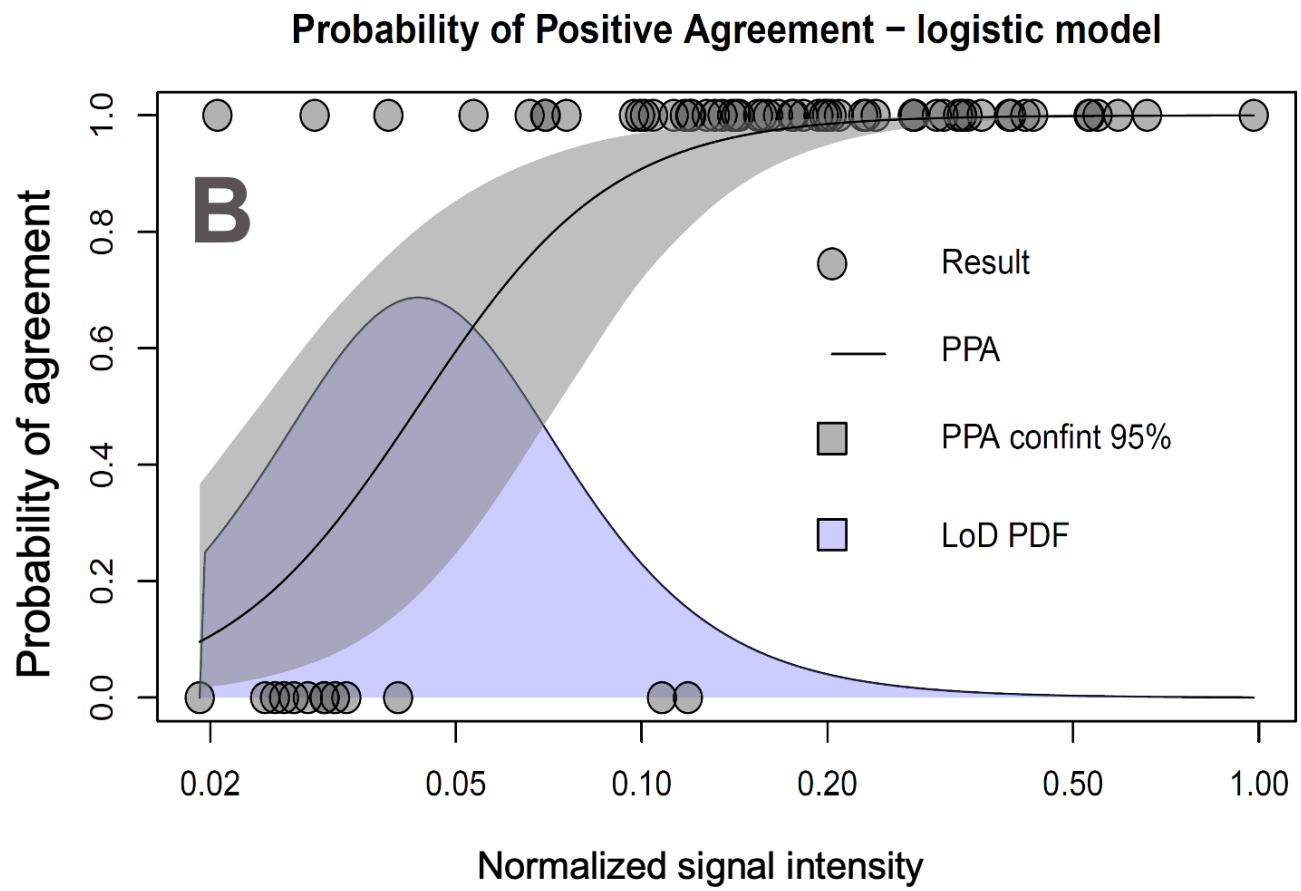
Figure 4. Characterization of the signal response with inactivated virus and domain-related transformation functions calibrated from experimental data. (A) The normalized signal response analysis for serial dilutions of inactivated virus. (B) Protein and virus concentration linear relational model based on the common signal intensity response of the devices, which allows transforming virus concentration to protein concentration and vice versa. (C) PCR cycle response calibration to inactivated virus dilution series. (D) qRT-PCR cycle response related to protein concentration, by combining transformations (B) and (C). LoD: limit of detection; PCR: polymerase chain reaction; pfu: plaque-forming unit; qRT-PCR: quantitative real-time polymerase chain reaction.



An additional component required by our predictive method is the LoD PDF of the observer's population pLoD(xlod) in the domain of the signal intensity. Figure 5A shows the real-world binary assessment conditioned to the signal intensity (based on Chelsea study participants' AT results and uploaded cell phone camera photos), the corresponding PPA function of the signal intensity estimated by logistic regression, and the LoD PDF in the domain of the signal intensity. The former is the observed LoD in the domain of the signal intensity for the participant population. Although, we can characterize the LoD in the domain of the signal intensity based on the real-world naked-eye data of the participant population, it is interesting to compare this estimation to one based on easier-to-obtain naked-eye data. Figure 5B shows naked-eye test results, the PPA function, and LoD PDF for trained staff assessing the AT result using a dilution series of recombinant SARS-CoV-2 nucleoprotein (including at a concentration of 0; ie, negatives) with blinded concentration labels. Acknowledging the nonnegligible effect of user heterogeneity [27] and external conditions, the probability functions (cumulative and density) were remarkably similar, as shown in Figures 5A and 5B.

Figure 5. Probability of positive agreement and LoD probability density for naked-eye assessments of antigen tests in common cassette presentation made by two different groups of observers. (A) Trained staff observing results from SARS-CoV-2 nucleoprotein dilutions blinded to the observer and (B) community participants of the study reporting self-tested results. In the case of trained staff, normalized intensity was calculated from laboratory environment photographs using cell phone cameras. The community participants' mobile phone photographs were uploaded to the digital reporting system and the visual assessment was a self-report also recorded in real time through the study's digital platform. LoD: limit of detection; PDF: probability density function.



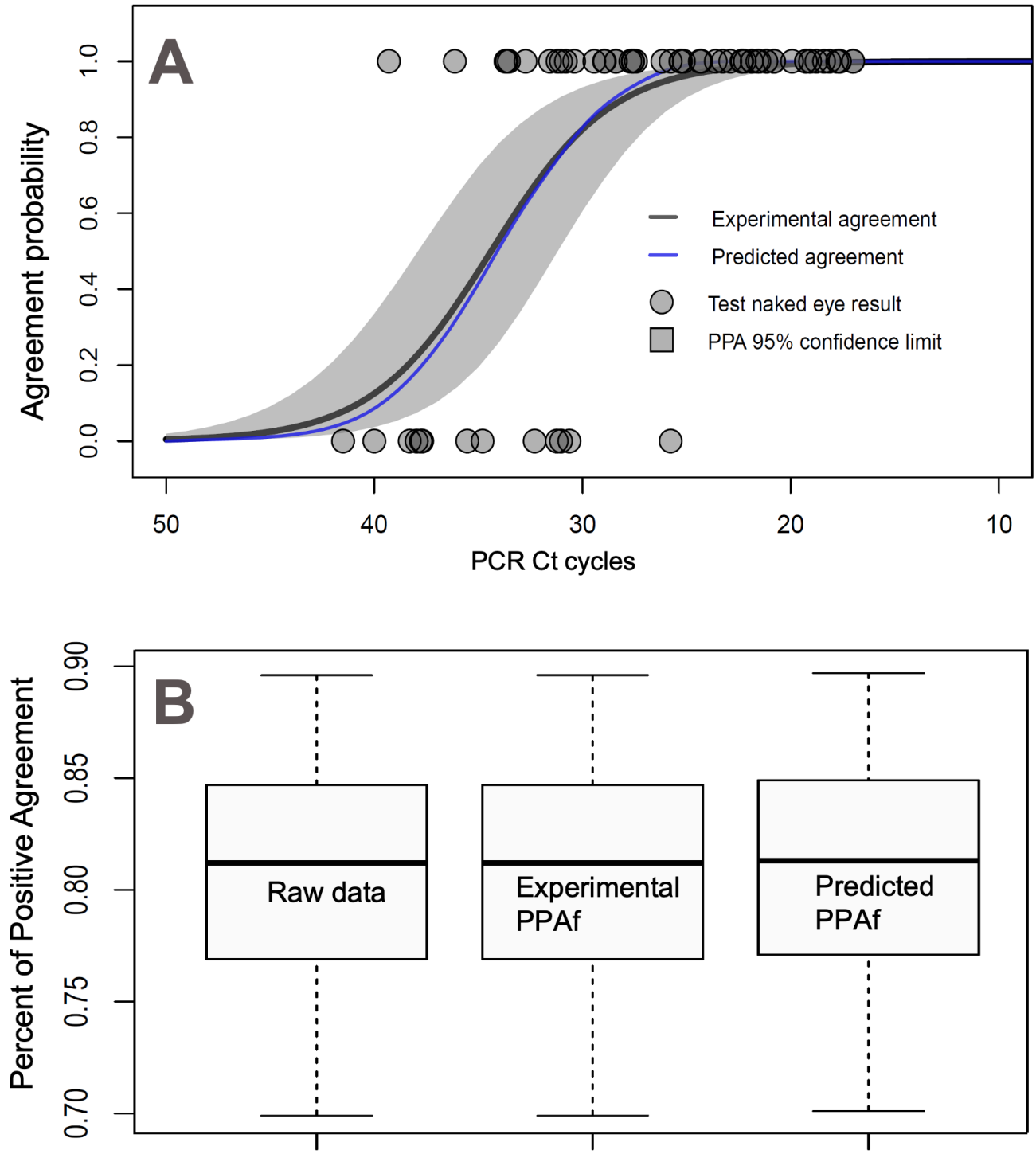


With the analysis shown in Figures 3-5, modeling the input components of the predictive model, we performed our

calculations of the PPA function of the qRT-PCR cycles and compared it with the *observed* PPA function (ie, derived from real-world data).

Figure 6A shows the real-world data of binary results obtained with the AT (ie, self-tested and logged data from the Chelsea study) and the corresponding observed PPA as a function of the qRT-PCR cycles. The calculation involved a description of the model uncertainties, illustrated in the plot with confidence intervals. Superimposed, we represented the predicted PPA function in the domain of the qRT-PCR, calculated using our method (Equation 8). The plot clearly shows that the predicted and observed PPA functions are very close and within the figure's confidence limits. Figure 6B shows a box plot of the overall data PPA for samples collected in the real-world setting for the AT and a comparison of the corresponding predicted real-world PPA according to expression 10. We can verify that both box plots provided close results, within statistical significance.

Figure 6. (A) Probability of agreement functions based on experimental real-world data (black curve) and resulting from our predictive model (blue curve). The experimental PPA function was obtained from a logistic regression analysis of the Chelsea project data for the test users; the gray area shows 95% confidence limits on the experimental PPA function. (B) Box plots showing the PPA calculated over the real-world raw data, the observed PPA function, and the predicted PPA function. Box plot bounds are Clopper-Pearson confidence limits for percentiles 50% (box) and 95% (segment). PPA: positive percent agreement; PCR: polymerase chain reaction.



Discussion

Principal Findings

Our predictive framework and the COVID-19 case study indicate that the real-world performance of ATs can be forecast from rapid laboratory measurements together with a statistical characterization of the population's LoD. First, analysis of test signal intensity and calibration to qRT-PCR can be completed entirely under controlled laboratory conditions—a rapid process that yields precise estimates of baseline AT performance. Second, we show that empirically characterizing the observer population's LoD distribution for visual interpretation of ATs is feasible.

Across two distinct observer groups and testing settings, the LoD PDFs and the resulting PPA–normalized signal intensity relationships were remarkably similar, suggesting a relatively

stable visual decision threshold inherent to this device design. We also demonstrated the practicality of paired AT/qRT-PCR sampling with image uploads, with minimal discrepancy between self-reported results and those recorded by trained staff. Nonetheless, environmental variability (lighting, optics and image compression, and differences among phone cameras), user motivation, and other external factors can broaden the effective LoD distribution.

Because we estimated the PPA as a continuous function of qRT-PCR Ct via logistic regression and then computed the expected sensitivity over a predefined reference Ct distribution, performance can be standardized across sites, time periods, and brands despite heterogeneous sampling. Bayesian fitting used in this study provides a practical path to hierarchical (brand-level) extensions we have recently reported in a complementary publication of the Chelsea clinical study.

Laboratory analyses and our predictive modeling framework provide a fast, standardized path to anticipate clinical performance and guide smarter evaluations of ATs, which is very useful information for the real-world validation of ATs. For regulatory decisions, quantitative summaries should be reported—PPA(Ct) with uncertainty bandwith and expected sensitivity under a preregistered reference Ct distribution—alongside observer LoD distributions (trained vs lay) and robustness to environmental and site effects, all with traceable calibrations and bridging analyses across lots, sites, and readers. After clearance, life cycle quality control should be maintained via routine lot verification, stability tracking, app-enabled field analytics for normalized signal intensity and invalid rates, anomaly detection, and timely label updates when variant-linked changes arise. At minimum, each evaluation should include the PPA(Ct) function with 95% CIs, expected sensitivity under a stated reference Ct distribution, observer LoD PDFs, calibration, and fit diagnostics.

The accurate calibration of the relational components (Figure 2) is fundamental for the fitness of the model. We performed triplicate dilution curves during the process of sample preparation for analysis. Additionally, we verified that thermal inactivation of the virus, as carried out for generating the serial virus dilutions, resulted in marginal loss of capsid protein detection (approximately a 2-fold difference), likely due to heat stability. We also used chemically inactivated virus alongside heat-inactivated virus stocks for signal intensity calibration.

The purpose of this study was to describe a method for the quantification and prediction of AT performance. Further work is underway, as we seek to expand our data for the comparative analysis of performance prediction across several test brands, comparative calibration of the qRT-PCR cycles across service providers, and further characterization of the signal intensity LoD for common ATs.

Conclusions

An accurate description of the AT signal intensity response conditioned by variables related to viral load, such as concentration of recombinant protein and concentration of inactivated virus, was established under laboratory conditions. These evaluations involved image processing of photographs and human naked-eye assessments using dilution series of the nucleoprotein of the SARS-CoV-2 virus. Modeling the performance on real testing conditions involved integrating the mentioned laboratory evaluation with information on the ability of the observer population to recognize the device's visual response, which can be described by the LoD PDF of the observer population in the domain of the test signal intensity. We described the overall test performance with the PPA *function* of the qRT-PCR Cts instead of using the common PPA (ie, sensitivity) for segments of clinical data. Our framework predicts PPA versus Ct by linking laboratory-normalized signal intensity–concentration models, observer LoD distributions, and a Ct–viral-load calibration. Conclusions are mathematical and are specific to the disease, assay, and setup. External validity depends on specific Ag-LFTs, user populations (for visual acuity and test interpretation), and qRT-PCR calibration across laboratories. Although the Chelsea dataset supports internal validity, broader clinical evaluation under intended-use conditions is required before generalized clinical claims, particularly for disease targets other than COVID-19. Therefore, the presented methodology has promising applications for the evaluation of ATs, as it involves a quick appraisal of real-world test performance.

Acknowledgments

We gratefully acknowledge the collaboration of Dr Alfred Harding and Dr Lee Gehrke for their virology expertise and for providing SARS-CoV-2 virus isolates to the study; Dr Karen Weeks and the personnel of Eco Laboratory, Acton, MA, for the processing and reporting of the nasal swab qRT-PCR data; Michael Fannon and BioIT Solutions personnel for their support on data management and reporting; Mateo Bonnet for front-end digital platform design and computer science expertise; the Center for Complex Intervention in Cambridge, MA, for community and scientific advice; the Chelsea Housing Authority; the Chelsea City Health Department; and the participants in this research for their utmost enthusiasm and support. We also extend our sincere thanks to the Reagan-Udall Foundation for the Food and Drug Administration personnel and their generous financial support through grant number 02282022 RUF, which enabled the development of this study. This research was supported by the Reagan-Udall Foundation for the Food and Drug Administration through grant number 02282022 RUF, awarded to IDX20 Inc.

Conflicts of Interest

IB is a founder of IDX20 Inc, a private company affiliated with this study. MB is a founder of Info Analytics Innovations, a private company affiliated with this study. The authors declare no additional conflicts of interest.

References

1. Wagenhäuser I, Knies K, Rauschenberger V, et al. Clinical performance evaluation of SARS-CoV-2 rapid antigen testing in point of care usage in comparison to RT-qPCR. EBioMedicine 2021 Jul;69:103455. [doi: 10.1016/j.ebiom.2021.103455] [Medline: 34186490]
2. Pickering S, Batra R, Merrick B, et al. Comparative performance of SARS-CoV-2 lateral flow antigen tests and association with detection of infectious virus in clinical specimens: a single-centre laboratory evaluation study. Lancet Microbe 2021 Sep;2(9):e461-e471. [doi: 10.1016/S2666-5247(21)00143-9] [Medline: 34226893]
3. Pollock NR, Savage TJ, Wardell H, et al. Correlation of SARS-CoV-2 nucleocapsid antigen and RNA concentrations in nasopharyngeal samples from children and adults using an ultrasensitive and quantitative antigen assay. J Clin Microbiol 2021 Mar 19;59(4):e03077-20. [doi: 10.1128/JCM.03077-20] [Medline: 33441395]
4. Karlafti E, Tsavdaris D, Kotzakioulafi E, et al. The diagnostic accuracy of SARS-CoV-2 nasal rapid antigen self-test: a systematic review and meta-analysis. Life (Basel) 2023 Jan 19;13(2):281. [doi: 10.3390/life13020281] [Medline: 36836639]
5. Wagenhäuser I, Knies K, Hofmann D, et al. Virus variant-specific clinical performance of SARS coronavirus two rapid antigen tests in point-of-care use, from November 2020 to January 2022. Clin Microbiol Infect 2023 Feb;29(2):225-232. [doi: 10.1016/j.cmi.2022.08.006] [Medline: 36028089]
6. Vaeth MJE, Abdullah O, Cheema M, et al. Accuracy of expired BinaxNOW rapid antigen tests. Microbiol Spectr 2023 Aug 17;11(4):e0208823. [doi: 10.1128/spectrum.02088-23] [Medline: 37428037]
7. Soni A, Herbert C, Lin H, et al. Performance of rapid antigen tests to detect symptomatic and asymptomatic SARS-CoV-2 infection: a prospective cohort study. Ann Intern Med 2023 Jul;176(7):975-982. [doi: 10.7326/M23-0385] [Medline: 37399548]
8. Corman VM, Haage VC, Bleicker T, et al. Comparison of seven commercial SARS-CoV-2 rapid point-of-care antigen tests: a single-centre laboratory evaluation study. Lancet Microbe 2021 Jul;2(7):e311-e319. [doi: 10.1016/S2666-5247(21)00056-2] [Medline: 33846704]
9. Bornemann L, Kaup O, Kleideiter J, et al. Virus variant-specific clinical performance of a SARS-CoV-2 rapid antigen test with focus on Omicron variants of concern. Clin Microbiol Infect 2023 Aug;29(8):1085. [doi: 10.1016/j.cmi.2023.05.009] [Medline: 37182639]
10. NIH study informs antigen testing for the SARS-cov-2 virus: repeat testing reduces false-negatives; FDA updates recommendations. National Institute of Biomedical Imaging and Bioengineering. 2025. URL: https://www.nibib.nih.gov/news-events/newsroom/nih-study-informs-antigen-testing-sars-cov-2-virus [accessed 2025-09-30]
11. Yoon CS, Park HY, Park HK, et al. The influence of pneumococcal positivity on clinical outcomes among patients hospitalized with COVID-19: a retrospective cohort study. PLoS One 2025;20(8):e0329474. [doi: 10.1371/journal.pone.0329474] [Medline: 40839694]
12. Testing for COVID-19. CDC. 2025. URL: https://www.cdc.gov/covid/testing/index.html [accessed 2025-09-30]
13. Kost GJ. The impact of increasing disease prevalence, false omissions, and diagnostic uncertainty on coronavirus disease 2019 (COVID-19) test performance. Arch Pathol Lab Med 2021 Jul 1;145(7):797-813. [doi: 10.5858/arpa.2020-0716-SA] [Medline: 33684204]
14. Stohr J, Wennekes M, van der Ent M, et al. Clinical performance and sample freeze-thaw stability of the cobas®6800 SARS-CoV-2 assay for the detection of SARS-CoV-2 in oro-/nasopharyngeal swabs and lower respiratory specimens. J Clin Virol 2020 Dec;133:104686. [doi: 10.1016/j.jcv.2020.104686] [Medline: 33221622]
15. Stohr J, Zwart VF, Goderski G, et al. Self-testing for the detection of SARS-CoV-2 infection with rapid antigen tests for people with suspected COVID-19 in the community. Clin Microbiol Infect 2022 May;28(5):695-700. [doi: 10.1016/j.cmi.2021.07.039] [Medline: 34363945]
16. Vaeth MJE, Cheema M, Omer S, et al. Self-administered versus clinician-performed BinaxNOW COVID rapid test: a comparison of accuracy. Microbiol Spectr 2024 Mar 5;12(3):e0252523. [doi: 10.1128/spectrum.02525-23] [Medline: 38349164]
17. Wagenhäuser I, Knies K, Pscheidl T, et al. SARS-CoV-2 antigen rapid detection tests: test performance during the COVID-19 pandemic and the impact of COVID-19 vaccination. EBioMedicine 2024 Nov;109:105394. [doi: 10.1016/j.ebiom.2024.105394] [Medline: 39388783]
18. Herbert C, Wang B, Lin H, et al. Performance of and severe acute respiratory syndrome coronavirus 2 diagnostics based on symptom onset and close contact exposure: an analysis from the Test Us at Home prospective cohort study. Open Forum Infect Dis 2024 Jun;11(6):ofae304. [doi: 10.1093/ofid/ofae304] [Medline: 38911947]
19. Harmon A, Chang C, Salcedo N, et al. Validation of an at-home direct antigen rapid test for COVID-19. JAMA Netw Open 2021 Aug 2;4(8):e2126931. [doi: 10.1001/jamanetworkopen.2021.26931] [Medline: 34448871]
20. Bosch I, de Puig H, Hiley M, et al. Rapid antigen tests for dengue virus serotypes and Zika virus in patient serum. Sci Transl Med 2017 Sep 27;9(409):eaan1589. [doi: 10.1126/scitranslmed.aan1589] [Medline: 28954927]
21. Safenkova I, Zherdev A, Dzantiev B. Factors influencing the detection limit of the lateral-flow sandwich immunoassay: a case study with potato virus X. Anal Bioanal Chem 2012 Jun;403(6):1595-1605. [doi: 10.1007/s00216-012-5985-8] [Medline: 22526658]
22. IDX20. 2025. URL: https://idx20.us/ [accessed 2025-09-12]

23. Douven S, Paez CA, Gommes CJ. The range of validity of sorption kinetic models. J Colloid Interface Sci 2015 Jun 15;448:437-450. [doi: 10.1016/j.jcis.2015.02.053] [Medline: 25765735]
24. Syafiuddin A, Salmiati S, Jonbi J, Fulazzaky MA. Application of the kinetic and isotherm models for better understanding of the behaviors of silver nanoparticles adsorption onto different adsorbents. J Environ Manage 2018 Jul 15;218:59-70. [doi: 10.1016/j.jenvman.2018.03.066] [Medline: 29665487]
25. Altın O, Özbelge H, Doğu T. Use of general purpose adsorption isotherms for heavy metal–clay mineral interactions. J Colloid Interface Sci 1998 Feb;198(1):130-140. [doi: 10.1006/jcis.1997.5246]
26. Saadi R, Saadi Z, Fazaeli R, Fard NE. Monolayer and multilayer adsorption isotherm models for sorption from aqueous media. Korean J Chem Eng 2015 May;32(5):787-799. [doi: 10.1007/s11814-015-0053-7]
27. De Nardo P, Tebon M, Savoldi A, et al. Diagnostic accuracy of a rapid SARS-CoV-2 antigen test among people experiencing homelessness: a prospective cohort and implementation study. Infect Dis Ther 2023 Apr;12(4):1073-1082. [doi: 10.1007/s40121-023-00787-0] [Medline: 36907951]

Abbreviations

Ag-LFT: antigen lateral flow test
AT: antigen test
Ct: cycle threshold
IRB: institutional review board
LoD: limit of detection
PCR: polymerase chain reaction
PDF: probability density function
PPA: percent positive agreement
qRT-PCR: quantitative real-time polymerase chain reaction
ROI: region of interest

Edited by F Wu; submitted 08.11.24; peer-reviewed by G Kost, HD Puig; revised version received 20.08.25; accepted 29.08.25; published 06.10.25.

Please cite as:
Bosch M, Garcia D, Rudtner L, Salcedo N, Colmenares R, Hoche S, Arocha J, Hall D, Moreno A, Bosch I
Real-World Performance of COVID-19 Antigen Tests: Predictive Modeling and Laboratory-Based Validation
JMIRx Med 2025;6:e68376
URL: https://xmed.jmir.org/2025/1/e68376
doi:10.2196/68376

Prevalence and Determinants of Academic Bullying Among Junior Doctors in Sierra Leone: Cross-Sectional Study

Fatima Jalloh[1], MB ChB; Ahmed Tejan Bah[2], MB ChB, MPH; Alieu Kanu[3], MB ChB; Mohamed Jan Jalloh[3], MB ChB; Kehinde Agboola[3], MB ChB; Monalisa M J Faulkner[3], MB ChB; Foray Mohamed Foray[4], MB ChB, MPH; Onome T Abiri[1], BPharm, PharmD, MSc; Arthur Sillah[5], PhD; Aiah Lebbie[1], MB ChB; Mohamed B Jalloh[6], MB ChB, MSc

[1]College of Medicine and Allied Health Sciences, University of Sierra Leone, Freetown, Sierra Leone

[2]Department of Public Health, Chamberlain College of Health Professions, Chicago, IL, United States

[3]University of Sierra Leone Teaching Hospitals Complex, Freetown, Sierra Leone

[4]College of Health Sciences and Public Policy, Walden University, Minneapolis, MN, United States

[5]School of Public Health, University of Washington, Seattle, WA, United States

[6]Faculty of Health Sciences, Department of Medicine, McMaster University, 1280 Main Street West, Hamilton, ON, Canada

Corresponding Author:
Mohamed B Jalloh, MB ChB, MSc
Faculty of Health Sciences, Department of Medicine, McMaster University, 1280 Main Street West, Hamilton, ON, Canada

Related Articles:

Companion article: https://www.medrxiv.org/content/10.1101/2024.11.13.24317261v1

Companion article: https://med.jmirx.org/2025/1/e75134

Companion article: https://med.jmirx.org/2025/1/e75135

Companion article: https://med.jmirx.org/2025/1/e75127

Abstract

Background: Academic bullying among junior doctors—characterized by repeated actions that undermine confidence, reputation, and career progression—is associated with adverse consequences for mental health and professional development.

Objective: This study aimed to investigate the prevalence and determinants of academic bullying among junior doctors in Sierra Leone.

Methods: We conducted a cross-sectional survey of 126 junior doctors at the University of Sierra Leone Teaching Hospitals Complex in Freetown between January 1 and March 30, 2024. Participants were selected through random sampling. Data were collected using a semistructured, self-administered questionnaire and analyzed with descriptive statistics and multivariable logistic regression.

Results: Of the 126 participants (n=77, 61.1% male; mean age 31.9, SD 5.05 years), 86 (68.3%) participants reported experiencing academic bullying. Among those, 55.8% (n=48) of participants experienced it occasionally and 36% (n=31) of participants experienced it very frequently. The most common forms were unfair criticism (n=63, 73.3%), verbal aggression (n=57, 66.3%), and derogatory remarks (n=41, 47.7%). Consultants and senior doctors were the main perpetrators, with incidents primarily occurring during ward rounds, clinical meetings, and academic seminars. No statistically significant predictors of bullying were found for gender (odds ratio 2.07, 95% CI 0.92 - 4.64; *P*=.08) or less than 2 years of practice (odds ratio 0.30, 95% CI 0.05 - 1.79; *P*=.19).

Conclusions: Academic bullying is widespread among junior doctors at the University of Sierra Leone Teaching Hospitals Complex. It has serious consequences for their mental health and professional development. There is an urgent need for clear and culturally appropriate policies, targeted training programs, confidential reporting systems, and leadership development. Promoting ethical leadership and fostering a culture of respect can help reduce incivility and burnout, leading to a healthier work environment for junior doctors.

(JMIRx Med 2025;6:e68865) doi:10.2196/68865

KEYWORDS

academic bullying; junior doctors; Sierra Leone; mental health; professional development

Introduction

Academic bullying—defined as maltreatment within academic settings intended to hinder the professional or academic progress of targeted individuals—remains a pervasive issue in medicine, particularly affecting junior doctors [1]. The hierarchical and high-stress nature of the medical profession, coupled with cultural norms that prioritize deference to authority, often creates environments in which public humiliation, verbal abuse, micromanagement, excessive workloads, and exclusion can flourish without adequate recourse [1-3]. Such repeated behaviors not only undermine the mental health and career development of junior doctors but also disrupt professional interactions and teamwork, potentially threatening patient safety and compromising the broader health care system [4].

Extensive research in high-income countries has consistently documented the widespread nature of bullying among junior doctors [1,4-6]. However, data from low-resource settings remain scarce. This paucity of information is especially concerning in Sierra Leone, where health care institutions grapple with significant resource limitations, workforce shortages, and constrained opportunities for professional development [7,8]. In these contexts, academic bullying may further intensify existing challenges, contributing to poor morale, reduced retention, and impaired patient care.

Adding to the urgency of investigating bullying within Sierra Leone's health care sector are studies that have already documented alarmingly high rates of bullying in the country's educational system. Research among in-school adolescents found a bullying prevalence of 48.7%, driven by factors such as loneliness, substance use, and school truancy [9]. School-related gender-based violence reports also confirm pervasive verbal and physical bullying, exacerbated by entrenched sociocultural norms and insufficient reporting mechanisms [10]. Although these data are drawn from younger populations, the same power imbalances and cultural drivers of bullying likely persist in higher education and professional settings. Indeed, the limited infrastructure for reporting and addressing maltreatment may allow such behaviors to continue into advanced academic and clinical environments.

Junior doctors in Sierra Leone are especially vulnerable to academic bullying due to strict hierarchies and limited resources, which can worsen their impact on both their well-being and the health care system. Despite the need for effective interventions, there is a lack of empirical data on the prevalence, determinants, and consequences of academic bullying among this demographic. Therefore, this study aims to investigate the prevalence of academic bullying among junior doctors at the University of Sierra Leone Teaching Hospitals Complex (USLTHC) in Freetown, Sierra Leone, and to examine the factors contributing to these behaviors. By situating the research within broader educational challenges and drawing on insights from prior studies, we seek to inform strategies for creating safer, more inclusive environments that support both the professional growth of junior doctors and the effective delivery of health care services.

Methods

Study Design and Setting

We conducted a cross-sectional survey at the major hospitals of USLTHC in Freetown, Sierra Leone. The USLTHC - Connaught Hospital, Princess Christian Maternity Hospital, Ola During Children's Hospital, and Sierra Leone Psychiatry Teaching Hospital are the largest and primary government referral hospitals in the country and serve as the main training centers for junior doctors, including registrars (residents) and house officers (interns). In Sierra Leone, the term "junior doctor" refers to physicians who have not yet achieved full specialist (consultant) status. This includes those in postgraduate training or supervised practice, such as house officers (interns), who are recent medical graduates undergoing closely supervised practice; medical officers, who have completed internships and can work more independently but have not pursued formal residency training; and registrars (residents), who are enrolled in specialty training programs but have not yet attained full accreditation as specialists. The survey was conducted from January 1, 2024, to March 30, 2024.

Participants and Sampling

All junior doctors who had been employed for a period of 6 months or longer and had reached the age of 18 years or older were included in the study. Those who were on outside posting or leave (annual or sick) were excluded, and no visiting junior doctors outside of USLTHC were included. The 6-month working experience requirement was used as the cutoff to ensure that participants have had sufficient interaction with both superiors and contemporaries during their training or postings.

Sampling Strategy and Sample Size

We constructed our sampling frame by compiling a list of all junior doctors aged 18 years or older who had been employed at the USLTHC for at least 6 months. From this roster, we used a computer-based random selection procedure (ie, assigning unique identifiers and using a random number generator) to ensure that each eligible junior doctor had an equal probability of inclusion. This method was chosen to maintain methodological rigor despite the logistical challenges posed by frequent 3- to 6-month rotations.

To determine the sample size for the study, we used the Yamane formula for cross-sectional studies: $n=N/(1+ N[e^2])$, where n is the required sample size, N is the total population size, and e is the margin of error set at 5% (0.05) [11].

Based on an estimated population of 160 eligible junior doctors, we calculated a minimum sample size of 114. Anticipating potential nonresponse or incomplete data, we increased this figure by 10% to arrive at a final target of 126 participants. Selected participants were drawn from multiple departments across four sites of the USLTHC.

Data Collection Instrument

Data were collected using a semistructured, self-administered questionnaire, also offered via web (via a secure server using Microsoft Forms) for participants who could not complete the paper-based version. The survey captured demographic details (eg, sex, age, duration of practice or training, and job title) and focused on first-hand encounters with workplace bullying within the preceding 6 months. Participants who reported bullying were asked to describe these incidents, ensuring the data represented direct, personal experiences rather than observations of others being bullied.

The primary outcome measure was the respondent's experience of workplace bullying, determined by a yes or no response to the question: "Have you experienced any form of workplace bullying in the last six months while training?"

Bullying was defined as repeated behaviors involving intimidation, humiliation, degradation, misuse of power, or abuse of authority that made the individual feel defenseless and undermined their dignity [1,2,12]. This definition guided our questionnaire design to differentiate self-reported experiences as a survivor from witnessing such acts. Before the main data collection, the survey instrument was piloted with 10 participants to confirm clarity and relevance, with refinements made based on their feedback.

Statistical Analysis

We conducted descriptive statistics to summarize the data. For continuous variables with a normal distribution, we reported means and SDs; for nonnormally distributed variables, medians and IQRs were provided. Associations between categorical variables were assessed using Pearson χ^2 tests or Fisher exact tests, as appropriate. Results were presented in tables and graphical summaries.

To explore independent associations between prespecified characteristics and the primary outcome—respondents' experience of workplace bullying—we performed multivariable logistic regression analyses. Explanatory variables were selected based on their relevance and included age (≤34 y vs ≥35 y), sex (male vs female), marital status (married vs others), level of training (house officer and others vs registrar), and duration of practice (≤2 y vs ≥3 y). Results were reported as odds ratios (ORs) with 95% CIs and corresponding *P* values. Statistical significance was set at a 5% level. All analyses were conducted using SPSS (version 27; IBM Corp).

Participant and Public Involvement Statement

Due to unexpected delays and time constraints, we were unable to involve participants or the public in the study's design, execution, or reporting. However, we are now considering a higher level of public and stakeholder engagement when sharing our research findings.

Ethical Considerations

The study received ethics approval from the College of Medicine and Allied Health Sciences Institutional Review Board (review number: COMAHS/IRB/013 - 2024). All procedures involving human participants were conducted in accordance with the ethical standards of the institutional and national research committee and with the 1964 Declaration of Helsinki and its later amendments. Informed consent was obtained from all participants prior to their completion of the questionnaire. Participation was voluntary, and participants were informed about the purpose of the study, their right to withdraw at any time, and the measures in place to protect their data. No compensation was provided to participants for their involvement in this study. All responses were collected anonymously, and no personally identifiable information was obtained. Strict confidentiality protocols were followed, including secure data storage and restricted access, to ensure the privacy and integrity of the data.

Results

Sociodemographic Characteristics of Participants

A total of 126 individuals completed the survey, comprising 77 (61.1%) male and 49 (38.9%) female participants. The mean age of the participants was 31.9 (SD 5.05) years. Regarding marital status, 68 (53.9%) individuals were single and never married, 52 (41.3%) individuals were married or in a domestic partnership, 2 (1.6%) individuals were separated, 1 (0.8%) individual was divorced, and 3 (2.4%) individuals preferred not to disclose their marital status (Table 1).

Table . Sociodemographic characteristics of the respondents (n=126).

Characteristics	Frequency
Age (years), n (%)	
18 - 24	2 (1.6)
25 - 34	95 (75.4)
35 - 44	26 (20.6)
45 - 54	3 (2.4)
Age (years), mean (SD)	31.9 (5.05)
Sex, n (%)	
Female	49 (38.9)
Male	77 (61.1)
Marital status, n (%)	
Single, never married	68 (53.9)
Married or domestic partnership	52 (41.3)
Separated	2 (1.6)
Divorced	1 (0.8)
Prefer not to say	3 (2.4)
Level of training, n (%)	
House officer	59 (46.8)
Medical officer	22 (17.5)
Registrar	43 (34.1)
Senior registrar	2 (1.6)
Duration of practice (years), n (%)	
<2	66 (52.4)
2 and above	60 (47.6)
Current training department, n (%)	
Internal medicine	35 (27.8)
Surgery and its subspecialties	34 (26.9)
Pediatrics	21 (16.7)
Obstetrics and gynecology	23 (18.3)
Family medicine	5 (3.9)
Psychiatry	6 (4.8)
Laboratory medicine	2 (1.6)

In terms of level of training, the sample included 59 (46.8%) house officers, 22 (17.5%) medical officers, 43 (34.1%) registrars, and 2 (1.6%) senior registrars. The duration of practice varied, with 66 (52.4%) participants having practiced for 2 years or less and 60 (47.6%) participants having practiced for 3 years or more (Table 1).

Participants were also categorized by their current training departments. Internal medicine had the highest representation, with 35 (27.8%) individuals, followed by surgery and its subspecialties with 34 (26.9%) individuals. Pediatrics included 21 (16.7%) participants, obstetrics and gynecology had 23 (18.3%) participants, family medicine included 5 (3.9%) participants, psychiatry had 6 (4.8%) participants, and laboratory medicine included 2 (1.6%) participants (Table 1).

This study examined the prevalence and forms of academic bullying among 126 participants. A total of 86 (68.3%) individuals reported experiencing bullying, while 40 (31.8%) individuals did not report such experiences (Table 2). Among the participants who reported being bullied, 48 (55.8%) experienced bullying occasionally, and more than one-third (36%) experienced bullying very frequently (Table 3).

Table . Prevalence and forms of academic bullying.

Variable	Frequency, n (%)
Current experience of bullying (n=126)	
Experienced	86 (68.3)
Not experienced	40 (31.8)
Forms of bullying (n=86)[a]	
Unfair criticism or evaluation	63 (73.3)
Verbal aggression	57 (66.3)
Derogatory remarks	41 (47.7)
Threat or intimidation	33 (38.4)
Undermining dignity at work	30 (34.9)
Exclusion from academic activities	16 (18.6)
Others (extra on-call service)	1 (1.2)
Common perpetrators of bullying (n=86)	
Consultants	72 (83.7)
Other senior doctors (colleagues)	66 (76.7)
Nursing staff	18 (20.9)
Administrative staff	15 (17.4)
Peers	13 (15.1)

[a]Percentages are calculated based on the total number of respondents who reported any form of bullying or reported any type of perpetrator (n=86).

Table . Frequency of bullying experienced by junior doctors.

Bullying frequency	Frequency, n (%)
Occasionally	48 (55.8)
Very frequently	31 (36.0)
Rarely	6 (7.0)
Always	3 (3.5)

Among those who reported experiencing bullying (n=86), the most common forms of bullying included unfair criticism or evaluation, reported by 63 (73.3%) individuals, and verbal aggression, reported by 57 (66.3%) individuals. Derogatory remarks were reported by 41 (47.7%) individuals, and threats or intimidation were experienced by 33 (38.4%) individuals. Other reported forms of bullying included undermining dignity at work (30/86 individuals, 34.9%), exclusion from academic activities (16/86 individuals, 18.6%), and extra on-call service demands (1/86 individuals, 1.2%) (Table 2).

Regarding the common perpetrators of bullying (n=86), consultants were identified as the most frequent perpetrators, reported by 72 (83.7%) individuals. Other senior doctors were reported by 66 (76.7%) individuals as perpetrators. Additionally, 18 (20.9%) individuals reported nursing staff as perpetrators, 15 (17.4%) individuals reported administrative staff, and 13 (15.1%) individuals reported peers as perpetrators of bullying (Table 2).

The most common context or setting in which academic bullying occurred was during ward rounds, reported by 73 (84.9%) participants. Clinical meetings were another context in which 51 (59.3%) individuals experienced bullying. A total of 50 (58.1%) individuals reported academic seminars or presentations as the context for bullying. Last, administrative meetings were identified as a bullying setting by 8 (9.3%) individuals (Table 4).

Table . Context or setting of bullying activity.

Context/setting	Frequency, n (%)
During ward rounds	73 (84.9)
Clinical meetings	51 (59.3)
Academic seminars or presentations	50 (58.1)
Administrative meetings	8 (9.3)

Multiple Logistic Regression Analysis of Factors Independently Associated With Bullying

The logistic regression analysis did not identify any statistically significant predictors of bullying at the 5% significance level. Participants aged 35 years or older had 0.78 times the odds of experiencing bullying compared with those aged 34 years or younger (OR 0.78, 95% CI 0.29-2.14; *P*=.63). House officers had 0.66 times the odds of experiencing bullying compared with registrars (OR 0.66, 95% CI 0.10-4.34; *P*=.67), while participants in the "Others" designation category (medical officers and senior registrars) had 2.58 times the odds of experiencing bullying compared with registrars (OR 2.58, 95% CI 0.67-9.92; *P*=.17). Marital status showed that participants categorized as "Others" had 0.94 times the odds of experiencing bullying compared with married or domestic partnership participants (OR 0.94, 95% CI 0.38-2.35; *P*=.90) (Table 5).

Table . Multiple logistic regression analysis of factors independently associated with bullying.

Factors	OR[a] (95% CI)	*P* value
Sex		
Female[b]	1	—[c]
Male	2.07 (0.92 - 4.64)	.08
Age (years)		
≤34[b]	1	—
35 or older	0.78 (0.29 - 2.14)	.63
Marital status		
Married or domestic partnership[b]	1	—
Others	0.94 (0.38 - 2.35)	.90
Level of training		
Registrar[b]	1	—
House officer	0.66 (0.10 - 4.34)	.67
Others	2.58 (0.67 - 9.92)	.17
Duration of practice (years)		
2 or more[b]	1	—
<2	0.30 (0.05 - 1.79)	.19
Intercept	3.00 (0.38 - 23.45)	.29

[a]OR: odds ratio.

[b]Reference categories that serve as the baseline for comparison.

[c]Not applicable.

Male participants had 2.07 times the odds of experiencing bullying compared with female participants (OR 2.07, 95% CI 0.92-4.64; *P*=.08). Participants with <2 years of practice had 0.30 times the odds of experiencing bullying compared with those with more than 2 years of practice (OR 0.30, 95% CI 0.05-1.79; *P*=.19) (Table 5).

The intercept, representing the log odds of experiencing bullying for the reference category (≤34 years old, female, married, registrar, ≥3 years of practice), had an OR of 3.00 (95% CI 0.38-23.45; *P*=.29), which serves as the baseline for comparison but is not directly interpretable in the same way as the other predictors (Table 5).

Discussion

Principal Findings

In this cross-sectional study, we investigated the prevalence and determinants of academic bullying among junior doctors at USLTHC in Freetown, Sierra Leone, between January 1 and March 30, 2024. We found a high prevalence of bullying (68.3%) among 126 participants, with unfair criticism and verbal aggression being the most common forms. Consultants and other senior doctors were frequently identified as perpetrators. Bullying occurred most frequently during ward rounds and clinical meetings. Despite the high prevalence, the analysis did not find any factors that were significantly associated with the likelihood of experiencing bullying.

The high prevalence of academic bullying in this study is much higher than the global average reported in systematic reviews, which found an overall prevalence of 51% (95% CI 36% - 66%) [4]. However, this finding aligns more closely with data from sub-Saharan Africa, exceeding the prevalence reported in Nigeria (59.7%) [2] but lower than that in Ghana (82%) [13]. These results suggest that while the prevalence of academic bullying in our study surpasses the global norm, it is consistent with regional trends.

Bullying predominantly occurred during ward rounds (84.9%), clinical meetings (59.3%), and academic seminars (58.1%), consistent with literature indicating that hierarchical settings in medical environments are common contexts for such behavior [14,15]. Multiple forms of bullying were identified, including unfair criticism, verbal aggression, derogatory remarks, and threats or intimidation. Consultants were the most frequently reported perpetrators, aligning with findings from a systematic review where 53.6% of 15,868 respondents identified senior staff as bullies [1]. These observations underscore the influence of entrenched power dynamics within the medical profession on bullying behaviors [16].

The high prevalence of bullying in our sample population can be attributed to several factors inherent in the medical profession. Hierarchical power dynamics, overwhelming workloads, and a lack of institutional support have been noted in other studies and are evident in our setting [14]. Bullying often occurs hierarchically, with senior staff perpetrating negative behaviors toward junior colleagues [15]. The Joint Commission has emphasized that health care professionals in positions of power commonly exhibit intimidating and disruptive behaviors, highlighting the systemic nature of the issue [16].

Toxic work cultures—including bullying and discrimination—are significant sources of distress for junior doctors, necessitating urgent institutional interventions. In Sierra Leone, medical professionals face escalating demands, diminishing resources, and staff shortages, factors known to compound psychological distress [7]. These stressors not only increase the risk of being bullied but also exacerbate the situations under which bullying occurs and intensify its negative impact. The absence of structured systems to counteract this culture may explain the high prevalence observed. Further research is needed to elucidate the role of these stressors, specifically related to perpetrators of bullying in the medical profession.

Determinants of Bullying in the Medical Profession

Our study found no significant differences in the incidence of bullying across demographic factors such as gender, age, marital status, designation, or duration of practice. While previous studies suggest a higher incidence of bullying against females [1,5]—and considering the patriarchal context of Sierra Leone—our data did not reflect significant gender differences. This may be due to reporting biases or specific workplace dynamics and aligns with findings from similar studies in the subregion [13,17]. These results underscore the need for further research and qualitative exploration to uncover underlying factors contributing to bullying.

Similarly, our findings deviate from other studies reporting higher odds of bullying among younger and less experienced individuals, attributed to lower status, perceived vulnerability, and power dynamics [18]. Studies have shown that individuals who are separated, divorced, or widowed have higher odds of reporting bullying than married individuals [19]. However, our study found no statistically significant correlation between marital status and reports of bullying.

The lack of statistically significant findings may be due to sample homogeneity; a more extensive and diverse sample could provide greater insight into demographic determinants of bullying, highlighting the need for further studies. Given the homogeneity of our sample, exploration of factors such as race-related bullying, which has been shown to lead to profound psychological distress, was not applicable [5].

Impact of Academic Bullying in the Medical Profession

Academic bullying has profound impacts on the medical profession. The hierarchical nature of medical training can lead to burnout and dissatisfaction among medical students and residents, deterring them from pursuing further specialization or academic careers [20]. This underscores the broader influence of workplace dynamics on health care professionals' career trajectories and well-being. In Sierra Leone, already facing a shortage of specialized medical staff, the negative effects of academic bullying may exacerbate this issue [7]. Research has demonstrated that victims of bullying may become perpetrators themselves, perpetuating a cycle particularly evident in hierarchical structures where each level may bully the one below [21].

Studies have highlighted the psychological impact of workplace bullying on junior doctors, including its associations with common mental disorders and suicidal ideation. The detrimental effects extend beyond direct victims to colleagues who may be vicariously impacted. Organizational factors, such as climate, culture, leadership, and support, play significant roles in predicting exposure to bullying, emphasizing the need for holistic approaches to address workplace victimization.

Research has also explored the relationship between workplace bullying and employee turnover intentions, as well as negative implications for productivity and teamwork [22]. The psychological and emotional distress caused by bullying affects both the personal and professional lives of junior doctors [23], a critical concern for nations like Sierra Leone grappling with medical professional shortages. While coping mechanisms such as seeking peer support and focusing on personal growth are used [24], systemic changes are imperative to address the root causes of bullying in academic settings. Recognizing workplace bullying as a systemic problem necessitates comprehensive solutions to foster a more supportive and respectful work environment.

Practical Implications

To effectively address academic bullying within USLTHC and the broader Sierra Leone health care system, a comprehensive, evidence-based approach is necessary. Establishing culturally sensitive antibullying policies is imperative to create a safer and more respectful academic environment. Implementing comprehensive training programs for medical staff—focused on recognizing and preventing bullying, promoting respectful communication, and fostering supportive work environments—is essential. Moreover, advocating for authentic leadership that empowers junior doctors, promotes transparent communication, and addresses hierarchical imbalances can substantially contribute to the mitigation of bullying behaviors in health care settings [25].

Confidential reporting channels, such as anonymous hotlines or independent web-based platforms, are vital for safeguarding individuals and promoting whistleblowing. Enhancing leadership development within the medical hierarchy is also crucial. Effective leadership models in health care enhance learning, teaching, and patient care. By fostering ethical leadership principles, health care organizations can cultivate a culture of respect, integrity, and accountability [26].

Ethical leadership profoundly influences health care outcomes, including job satisfaction, safety compliance, and reduction of workplace deviance. The positive impact of ethical leadership on job satisfaction enhances service quality, patient satisfaction, and productivity [27]. Ethical leadership improves safety compliance by building trust among health care professionals [28]. Fostering a culture of trust and ethical behavior is therefore crucial for promoting positive outcomes in health care organizations.

Addressing incivility and unethical behaviors in health care settings is essential. Organizations can leverage Ethics Committees and Clinical Ethics Consultation Services to manage incivility and promote ethical practices [29]. Integrating ethical considerations into organizational practices fosters a supportive and respectful work environment, aligning with the need to cultivate ethical leadership skills among health care professionals [30].

Implementing antibullying interventions and creating supportive environments through mentorship, coaching, and feedback mechanisms can mitigate the negative impacts of bullying on junior doctors [31,32]. Fostering a culture of respect and support within medical institutions is essential to promoting the well-being and professional development of all health care professionals, including junior doctors [20,33].

Strengths and Limitations

This study represents the first investigation into academic bullying among junior doctors in Sierra Leone. Strengths include the straightforward administration of the survey, facilitated by a well-educated study population and a readily accessible participant list.

However, several limitations must be acknowledged. The reliance on self-reported experiences introduces the potential for response bias, including underreporting due to fear of administrative scrutiny. Additionally, there is a lack of a validated instrument for evaluating academic bullying in an African context. The questionnaire was developed based on prior studies and an extensive literature review. Despite these constraints, the findings suggest disturbingly high levels of perceived bullying and mistreatment during training. Results should be interpreted cautiously, and a higher response rate would have been preferable.

Conclusions

This study revealed a high prevalence of academic bullying among junior doctors at USLTHC, with unfair criticism, verbal aggression, derogatory remarks, and threats or intimidation being the most common forms identified. Consultants and other senior doctors were frequently identified as perpetrators. Bullying most commonly occurs during ward rounds and clinical meetings. Despite the high prevalence, the analysis did not find any sociodemographic factors significantly associated with the likelihood of experiencing bullying.

Academic bullying in medicine undermines junior doctors' mental health and professional development, compromising both individual well-being and the quality of patient care. Confronting this pervasive issue within USLTHC and the broader Sierra Leone health care system demands a comprehensive, evidence-based strategy.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

All authors were involved in the conceptualization and planning of the study. FJ, AK, MJJ, KA, MMJF, and MBJ were involved in conducting the study, with data collection. FJ, AL, and MBJ were involved with the analysis and interpretation of data. FJ and MBJ prepared the first draft of the manuscript. All authors contributed to subsequent revisions to the manuscript. All authors read and approved the final manuscript.

Conflicts of Interest

None declared.

References

1. Averbuch T, Eliya Y, Van Spall HGC. Systematic review of academic bullying in medical settings: dynamics and consequences. BMJ Open 2021 Jul 12;11(7):e043256. [doi: 10.1136/bmjopen-2020-043256] [Medline: 34253657]
2. Afolaranmi TO, Hassan ZI, Gokir BM, et al. Workplace bullying and its associated factors among medical doctors in residency training in a tertiary health institution in plateau state Nigeria. Front Public Health 2021 Jan 27;9(January):812979. [doi: 10.3389/fpubh.2021.812979] [Medline: 35155359]
3. Argo Widiharto C, Kusdaryani W, Fitriana S, Prasetyo A. Academic resilience of students who are bullied. KnE Soc Sci 2022. [doi: 10.18502/kss.v7i19.12449]

4. Álvarez Villalobos NA, De León Gutiérrez H, Ruiz Hernandez FG, Elizondo Omaña GG, Vaquera Alfaro HA, Carranza Guzmán FJ. Prevalence and associated factors of bullying in medical residents: a systematic review and meta-analysis. J Occup Health 2023;65(1):e12418. [doi: 10.1002/1348-9585.12418] [Medline: 37443455]
5. Camm CF, Joshi A, Moore A, et al. Bullying in UK cardiology: a systemic problem requiring systemic solutions. Heart 2022 Feb;108(3):212-218. [doi: 10.1136/heartjnl-2021-319882] [Medline: 34872975]
6. Iyer MS, Way DP, MacDowell DJ, Overholser BM, Spector ND, Jagsi R. Bullying in academic medicine: experiences of women physician leaders. Acad Med 2023 Feb 1;98(2):255-263. [doi: 10.1097/ACM.0000000000005003] [Medline: 36484542]
7. Jalloh MB, Naveed A, Johnson SAA, et al. Perceptions of burnout among public sector physicians in Sierra Leone: a qualitative study. PLOS Glob Public Health 2024 Sep 16;4(9):e0003739. [doi: 10.1371/journal.pgph.0003739] [Medline: 39283876]
8. Johnson O, Sahr F, Sevdalis N, Kelly AH. Exit, voice or neglect: understanding the choices faced by doctors experiencing barriers to leading health system change through the case of Sierra Leone. SSM Qual Res Health 2022 Dec 2;2(May):100123. [doi: 10.1016/j.ssmqr.2022.100123] [Medline: 36531296]
9. Osborne A, James PB, Bangura C, Tom Williams SM, Kangbai JB, Lebbie A. Bullying victimization among in-school adolescents in Sierra Leone: a cross-sectional analysis of the 2017 Sierra Leone global school-based health survey. PLOS Glob Public Health 2023 Dec 22;3(12):e0002498. [doi: 10.1371/journal.pgph.0002498] [Medline: 38134001]
10. Report on findings from school-related gender-based violence action research in schools and communities in sierra leone. UNGEI.: UNICEF; 2023 Jan. URL: https://www.ungei.org/publication/report-findings-school-related-gender-based-violence-action-research-schools-and [accessed 2025-05-16]
11. Yamane T. Statistics: An Introductory Analysis (A Harper International Edition): Harper & Row; 1967. URL: https://books.google.ca/books?id=Wr7rAAAAMAAJ [accessed 2025-05-16]
12. Rajalakshmi M, Gomathi S. A study on the factors influencing workplace bullying and its impact on employee stress. Mediterranean J Soc Sci 2015 Jan 1;6(1):292-299. [doi: 10.5901/mjss.2015.v6n1p292]
13. Anyomih TTK, Mehta A, Wondoh PM, Mehta A, Siokos A, Adjeso T. Bullying among medical students and doctors in Ghana: a cross-sectional survey. Singapore Med J 2024 Apr 29. [doi: 10.4103/singaporemedj.SMJ-2021-281] [Medline: 38779930]
14. Leisy HB, Ahmad M. Altering workplace attitudes for resident education (A.W.A.R.E.): discovering solutions for medical resident bullying through literature review. BMC Med Educ 2016 Apr 27;16:127. [doi: 10.1186/s12909-016-0639-8] [Medline: 27117063]
15. Hussain NM, Spiers J, Kobab F, Riley R. The impact of race and gender-related discrimination on the psychological distress experienced by junior doctors in the UK: a qualitative secondary data analysis. Healthcare (Basel) 2023 Mar 12;11(6):834. [doi: 10.3390/healthcare11060834] [Medline: 36981491]
16. Riley R, Buszewicz M, Kokab F, et al. Sources of work-related psychological distress experienced by UK-wide foundation and junior doctors: a qualitative study. BMJ Open 2021 Jun 23;11(6):e043521. [doi: 10.1136/bmjopen-2020-043521] [Medline: 34162634]
17. Darko G, Björkqvist K, Österman K. Workplace bullying and psychological distress in public institutions in Ghana. Eur J Soc Sci Educ Res 2019;6(1):62. [doi: 10.26417/ejser.v6i1.p62-74]
18. Hayat A, Afshari L. Supportive organizational climate: a moderated mediation model of workplace bullying and employee well-being. Pers Rev 2021 Oct 17;50(7/8):1685-1704. [doi: 10.1108/PR-06-2020-0407]
19. Keuskamp D, Ziersch AM, Baum FE, Lamontagne AD. Workplace bullying a risk for permanent employees. Aust N Z J Public Health 2012 Apr;36(2):116-119. [doi: 10.1111/j.1753-6405.2011.00780.x] [Medline: 22487344]
20. Samsudin EZ, Isahak M, Rampal S, Rosnah I, Zakaria MI. Individual antecedents of workplace victimisation: the role of negative affect, personality and self-esteem in junior doctors' exposure to bullying at work. Int J Health Plann Manage 2020 Sep;35(5):1065-1082. [doi: 10.1002/hpm.2985] [Medline: 32468617]
21. Hauge LJ, Skogstad A, Einarsen S. Individual and situational predictors of workplace bullying: why do perpetrators engage in the bullying of others? Work Stress 2009 Oct;23(4):349-358. [doi: 10.1080/02678370903395568]
22. Otema OD, Acanga AA, Mwesigwa DM. Workplace bullying and its consequence to employee productivity in civil society organisations in Lira City, Uganda. Hum Resource Leadership J 2022 Dec 28;7(2):95-108. [doi: 10.47941/hrlj.1159]
23. Dunning A, Teoh K, Martin J, et al. Relationship between working conditions and psychological distress experienced by junior doctors in the UK during the COVID-19 pandemic: a cross-sectional survey study. BMJ Open 2022 Aug 23;12(8):e061331. [doi: 10.1136/bmjopen-2022-061331] [Medline: 35998957]
24. Howe A, Smajdor A, Stöckl A. Towards an understanding of resilience and its relevance to medical training. Med Educ 2012 Apr;46(4):349-356. [doi: 10.1111/j.1365-2923.2011.04188.x] [Medline: 22429170]
25. Malila N, Lunkka N, Suhonen M. Authentic leadership in healthcare: a scoping review. Leadersh Health Serv (Bradf Engl) 2018 Feb 7;31(1):129-146. [doi: 10.1108/LHS-02-2017-0007] [Medline: 29412093]
26. Hargett CW, Doty JP, Hauck JN, et al. Developing a model for effective leadership in healthcare: a concept mapping approach. J Healthc Leadersh 2017 Aug 28;9:69-78. [doi: 10.2147/JHL.S141664] [Medline: 29355249]

27. Ahmad I, Umrani WA. The impact of ethical leadership style on job satisfaction. Leadersh Organ Dev J 2019 Jul 8;40(5):534-547. [doi: 10.1108/LODJ-12-2018-0461]
28. Enwereuzor IK, Adeyemi BA, Onyishi IE. Trust in leader as a pathway between ethical leadership and safety compliance. Leadership Health Service 2020 Mar 13;33(2):201-219. [doi: 10.1108/LHS-09-2019-0063]
29. Blackler L, Scharf AE, Chin M, Voigt LP. Is there a role for ethics in addressing healthcare incivility? Nurs Ethics 2022 Sep;29(6):1466-1475. [doi: 10.1177/09697330221105630] [Medline: 35724428]
30. Sakr F, Haddad C, Zeenny RM, et al. Work ethics and ethical attitudes among healthcare professionals: the role of leadership skills in determining ethics construct and professional behaviors. Healthcare (Basel) 2022 Jul 27;10(8):1399. [doi: 10.3390/healthcare10081399] [Medline: 35893220]
31. Ling M, Young CJ, Shepherd HL, Mak C, Saw RPM. Workplace bullying in surgery. World J Surg 2016 Nov;40(11):2560-2566. [doi: 10.1007/s00268-016-3642-7] [Medline: 27624759]
32. de Lasson L, Just E, Stegeager N, Malling B. Professional identity formation in the transition from medical school to working life: a qualitative study of group-coaching courses for junior doctors. BMC Med Educ 2016 Jun 24;16:165. [doi: 10.1186/s12909-016-0684-3] [Medline: 27342973]
33. Abdelaziz EM, Abu-Snieneh HM. The impact of bullying on the mental health and academic achievement of nursing students. Perspect Psychiatr Care 2022 Apr;58(2):623-634. [doi: 10.1111/ppc.12826] [Medline: 33949687]

Abbreviations

OR: odds ratio
USLTHC: University of Sierra Leone Teaching Hospitals Complex

Edited by S Tungjitviboonkun; submitted 15.11.24; peer-reviewed by J Wilkinson, PB James; revised version received 14.03.25; accepted 15.03.25; published 22.05.25.

Please cite as:
Jalloh F, Bah AT, Kanu A, Jalloh MJ, Agboola K, Faulkner MMJ, Foray FM, Abiri OT, Sillah A, Lebbie A, Jalloh MB
Prevalence and Determinants of Academic Bullying Among Junior Doctors in Sierra Leone: Cross-Sectional Study
JMIRx Med 2025;6:e68865
URL: https://xmed.jmir.org/2025/1/e68865
doi:10.2196/68865

The Impact of Rural Alimentation on the Motivation and Retention of Indigenous Community Health Workers in India: A Qualitative Study

Ajit Kerketta*, MHA; Raghavendra A N*, PhD

CHRIST (Deemed to be University), Hosur Road, Bhavani Nagar, Bengaluru, India

*all authors contributed equally

Corresponding Author:
Ajit Kerketta, MHA
CHRIST (Deemed to be University), Hosur Road, Bhavani Nagar, Bengaluru, India

Related Articles:

Companion article: https://www.medrxiv.org/content/10.1101/2023.04.12.23288461v1

Companion article: https://med.jmirx.org/2025/1/e70808

Companion article: https://med.jmirx.org/2025/1/e70059

Abstract

Background: Rural health care delivery remains a global challenge and India is no exception, particularly in regions with Indigenous populations such as the state of Jharkhand. The Community Health Centres in Jharkhand, India, are staffed by Indigenous workers who play a crucial role in bridging the health care gap. However, their motivation and retention in these challenging areas are often influenced by a complex mix of sociocultural and environmental factors. One such significant but understudied influencing factor is alimentation, or nutrition, in rural settings. Previous studies have identified several motivators, including community ties, cultural alignment, job satisfaction, and financial incentives. However, the role of alimentation in their motivation and retention in rural areas has not been sufficiently explored.

Objective: This study aims to explore how the strong bond with locally produced food products impacts the retention of Indigenous community health workers (CHWs) in Jharkhand, India, and shed light on a crucial aspect of rural health care workforce sustainability.

Methods: This study adopted a phenomenological research design to explore the lived experiences and perspectives of Indigenous CHWs in Jharkhand. A purposive sampling method was used to select CHWs who had worked in rural areas for at least five years. Data were collected through semistructured interviews, focusing on the participants' experiences of rural alimentation and how it influences their motivation and retention for rural health care. The interviews were audio recorded, transcribed, and analyzed using thematic analysis to identify common themes and patterns in their experiences related to nutrition and retention.

Results: The study revealed that rural alimentation plays a significant role in both the motivation and retention of CHWs in Jharkhand. CHWs who experienced consistent access to local food reported higher job satisfaction, better physical well-being, and a stronger commitment to their roles. It has also been perceived that consuming nutrient-dense food products decreases the risk of chronic illness among rural populations. Additionally, community support systems related to alimentation were found to be crucial in maintaining motivation, with many CHWs emphasizing the importance of local food availability and cultural ties. The findings suggest that improving access to organic nutrition can positively influence the retention of CHWs in rural areas.

Conclusions: Indigenous communities have unique food habits and preferences deeply rooted in agriculture and arboriculture. Their traditional eating practices are integral to their rich cultural heritage, with significant social, symbolic, and spiritual importance. This study highlights the critical role of rural alimentation in motivating and retaining CHWs in rural Community Health Centres in Jharkhand. Therefore, addressing organic versus conventional food in rural health care policies plays a vital role in improving the retention rates of CHWs. By recognizing the interconnectedness of nutrition and workforce sustainability, health care systems can better support Indigenous CHWs and continue delivering health care services.

(JMIRx Med 2025;6:e48346) doi:10.2196/48346

KEYWORDS

rural alimentation; community health workers; motivation; retention; rural health; rural nutrition; workforce

Introduction

Rural Health and Alimentation

Community health workers (CHWs) play a vital role in providing primary health care services to rural populations in low- and middle-income countries [1,2]. However, retaining them in rural areas is challenging, largely due to low motivation. One potential factor influencing their motivation and retention is access to a diverse and nutritious diet or rural alimentation [3]. Although the term "alimentation" has existed in the English language since the late 16th century, it is rarely used. In Latin-based languages like French, "alimentation" conveys a holistic view of how humans produce, procure, prepare, share, consume, and digest their food, encompassing human, technological, sociocultural, and environmental aspects [4].

Significance of Alimentation in Rural Health Systems

The term "rural alimentation" in this study refers to the food that Indigenous people produce, acquire, prepare, share, consume, and digest; it is intimately linked to their sociocultural and environmental surroundings. Indigenous CHWs are also among those who are devoted to these tastes and preferences and will find it difficult to give up their native cuisine. Rural food products that are fresh, pure, unadulterated, nutrient dense, and low in pesticides appeal to CHWs [5]. This experience has lured CHWs to continue serving in rural health centers in Jharkhand, India. However, the lack of local and traditional food in metropolitan cities has negatively impacted their motivation, causing many health workers to select rural health care jobs [6]. Previous studies indicate that the desire for nutritious food in the urban setting significantly affects their motivation, job satisfaction, and retention rates. For example, a study in Ethiopia showed that providing nutritious food to CHWs increased job satisfaction and reduced attrition rates [7]. Similarly, a study conducted in Malawi demonstrated that CHWs accessing local food products were less likely to leave their jobs [8]. Despite the potential impact of rural alimentation on CHWs' motivation and retention, there is limited research on this topic in Jharkhand. Therefore, the study seeks to explore the question: "How does access to diverse and organic food in rural Jharkhand influence the motivation and retention of Indigenous community health workers?"

The study also aims to explore how Indigenous CHWs in Jharkhand perceive the impact of rural alimentation on their motivation and retention. Investigating the connection between rural alimentation and the motivation and retention of Indigenous CHWs will provide valuable insights into the factors that influence their engagement and commitment. We also intend to share the findings with policy makers and health care stakeholders, invoking the implementation of policies that support the well-being of CHWs, promote local food, and attract adequate CHWs to rural Jharkhand.

Study Background

Jharkhand is an eastern Indian state with a population of 39 million spread across 79,714 square kilometers (2019 census). Out of the total population, 24.05% reside in urban areas, while 75.95% live in rural areas [9]. Agriculture and agroforestry products are the primary sources of livelihood. However, the younger generation is increasingly migrating to metropolitan cities in search of better, more sustainable living opportunities. Most state regions are characterized by hills, rugged terrain, lakes, and rivers, presenting significant challenges. While some areas have plains and level topography nestled within natural surroundings, socioeconomic difficulties and a lack of infrastructure make it challenging for CHWs to stay in these locations. Consequently, Jharkhand faces a severe shortage of health workforce [10,11]. Approximately 80% of health care workers are stationed in metropolitan cities catering to the 24.05% of the population residing in urban areas, while 20% health care workers serve the 75.95% population living in rural areas [11-13]. Additionally, the population's strong beliefs in spirit worship and reliance on local quacks and tantric practices for their ill health further contribute to the short supply of CHWs. This study aims to provide evidence-based insights into the factors that promote the retention of CHWs in rural areas.

Methods

Study Design

We used a qualitative case research design to help understand the perspectives, emotions, and behaviors of Indigenous CHWs and uncover their in-depth experiences [14]. This approach focuses on understanding the subjective meaning that drove CHWs to work in rural Jharkhand [15,16]. This study selected participants with 5 years of service records in the respective Community Health Centers (CHCs).

Ethical Considerations

The study is not a clinical trial and, therefore, does not require registration to establish safety and efficacy standards. Nevertheless, ethical approval was obtained from the Institutional Review Board of CHRIST (Deemed to be University), Bangalore, India (CU: RCEC/00371/11/22). Written informed consent was obtained from each participant before conducting the interview. To further ensure the privacy of the participants, all names were changed to pseudonyms during the transcriptions of the text. However, the interviewers (AK and RNA) know the actual names of the interview participants. Each participant received a fixed remuneration of US $5.75 after completion of the interview as an acknowledgment of their time and contribution.

Setting and Sample

The corresponding author randomly selected and visited 3 CHCs in the eastern districts of Jharkhand to pilot the survey. This visit played a key role in shaping the development of the research objective: to explore the impact of rural alimentation (local and traditional food systems) on the motivation and

retention of Indigenous CHWs in rural India, as well as to establish a suitable research framework. CHWs from these randomly selected CHCs participated by completing a self-validated questionnaire with open-ended questions. In the main study, 30 CHWs were selected; of these, 10, 12, and 8 CHWs from the respective CHCs met the study criteria. They had served more than five years, expressed willingness to continue residing in rural areas, and were government employees. The study adopted a purposive sampling technique, which helped obtain rich, detailed, and relevant data that influenced the motivation and retention of CHWs in Jharkhand [17]. The male and female respondents were selected irrespective of their rural and urban backgrounds.

Process of Data Generation

A total of 14 participants (4 male individuals and 10 female individuals) ultimately consented to participate in the interviews. However, 16 individuals declined, with some initially agreeing but later withdrawing due to hesitation from the novelty of such an interview process and discomfort with having their comments audio recorded. The participants were aged 30-60 years and expressed their desire to participate in individual, face-to-face or telephone interviews within 10 months. A follow-up interview was done after 4 and 6 months. Within 4 months, 8 interviews were conducted at the CHCs and 6 interviews were conducted in the home district of the reviewer [18,19]. Multimedia Appendix 1 shows the interview guidelines and questionnaire.

The round-1 interview was precise and relevant to the objectives mentioned above (in the *Setting and Sample* section) and, hence, did not require a reinterview of any participants. Interviews were conducted both face-to-face and remotely in Hindi, a language in which the authors are fluent and experienced in conducting qualitative case research. While consent was sought to audio record the interviews, many participants expressed unwillingness; as a result, the researchers took detailed notes instead.

The data were collected through individual, semistructured qualitative case research, with in-depth interviews conducted according to the established protocol matrix [20]. Questions regarding all main areas were posed, albeit in varying order. The interviews in the 4-month follow-up ranged between 6 and 37 minutes (average of 10 min), and interviews in the 6-month follow-up ranged between 5 and 13 minutes (average of 7 min).

Research Team and Reflexibility

AK (research scholar in human resource management, male, aged 40 y) and RAN (PhD in human resource management, male, aged 48 y) solely conducted the interviews. After the interviews, the corresponding author listened to the audio recordings, with several breaks between every audio recording, and transcribed them.

Analysis

We employed the general data analysis methodologies indicated below in the context of thematic analysis and read the texts multiple times to familiarize and better understand them [21]. Descriptive codes were then applied to data segments [22] relevant to the research question: how do the local food habits influence motivation and continuation of work, and do these factors impact decisions to remain in rural areas? This question was aligned with the objective of the study [23]. The coded data were grouped into themes using QDA Miner Lite software (Provalis Research), demonstrating the relationships between them and identifying themes using inductive methods. The themes were assessed and modified depending on their relevance to the data and the research topic, and they were blended as appropriate. After the themes were developed, they were further defined and given titles that accurately expressed their meanings [24]. Then, the researcher drafted the report. The thematic analysis involves a recursive process of moving back and forth between the data and the emerging themes. It is an iterative and reflexive process, requiring the researcher to consider their biases and assumptions throughout the analysis.

- In-depth investigation: This method provided an in-depth understanding of the study's objectives and phenomena [25]. It enabled the researchers to collect data from multiple sources and examine them comprehensively.
- Contextual analysis: The qualitative case research design allowed the researchers to focus on the social, cultural, economic, and political factors influencing the phenomenon [26].
- Interpretive analysis: This approach involved identifying themes and interpreting them in the context of the research objectives [21,27].
- Flexible design: The qualitative case research design is adaptable, allowing the researchers to evolve the design as data are collected and analyzed [28]. To explore complex and context-specific issues in real-life settings, the interview provided comprehensive insights into the CHWs' experiences, opinions, and perspectives regarding rural alimentation and its impact on their motivation and retention.

Results

Study Participants

We contacted 4 CHCs; however, the medical officer at one center declined to grant permission, citing concerns that the study might inadvertently violate government protocols. A total of 64 CHWs were contacted across the remaining 3 CHCs, who were directly appointed by the government and were under the age of 60 years. Table 1 shows that the majority of health workers were female, accounting for 52% (13/25), 61% (11/18), and 57% (12/21) across the 3 CHCs. Among the 14 participants, 71% (n=10) were female and 29% (n=4) were male. This sex disparity could be a potential area for further research, exploring why fewer male CHWs tend to remain in rural locations.

Table . Participants characteristics from CHCs[a] A, B, and C. This table combines the demographic and workplace preferences of health care workers across the 3 centers (A, B, and C).

Characteristics	Center A (n=25), n (%)	Center B (n=18), n (%)	Center (n=21), n (%)
Sex			
Male	12 (48)	7 (39)	9 (43)
Female	13 (52)	11 (61)	12 (57)
Age group (years)			
≤30	3 (12)	5 (28)	4 (19)
≥30	22 (88)	13 (72)	17 (81)
Residence			
Rural origin	25 (100)	18 (100)	21 (100)
Urban origin	0 (0)	0 (0)	0 (0)
Preferred workplace			
Rural area	8 (32)	12 (67)	15 (71)
Male	2 (25)[b]	4 (33)[b]	4 (27)[b]
Female	6 (75)[b]	9 (67)[b]	11 (73)[b]
Urban area	17 (68)	6 (33)	6 (29)
Male	3 (18)[c]	1 (17)[c]	2 (33)[c]
Female	14 (82)[c]	5 (83)[c]	4 (67)[c]

[a]CHC: Community Health Centre.

[b]Percentages are based the number of workers who preferred a rural workplace as the denominator.

[c]Percentages are based the number of workers who preferred an urban workplace as the denominator.

Data were analyzed by constructing a thematic analysis, identifying patterns and themes as guided by the research questions and objectives [24,29]. Emerging themes were verified through member checking to ensure accuracy and validity. This study offers a comprehensive understanding and valid representations [30] of the perspectives and experiences of CHWs staying in rural Jharkhand. The focus is on a specific area within the CHCs, which is predominantly tribal dominated. The analysis identified themes that offered insights into the barriers and facilitators affecting CHWs' access to and consumption of diverse and nutritious food, as well as how their food habits intersect with their roles as health promoters and caregivers.

The study explored three major themes, presented as main themes and their corresponding minor themes, as illustrated below. These themes reflect the perspectives, experiences, and perceptions of the Indigenous CHWs regarding their reasons for remaining in rural Jharkhand.

1. The impact of rural alimentation on Indigenous CHWs' motivation
2. Retention trends among Indigenous CHWs
3. Correlations between nutritional support and job satisfaction

Impact of Rural Alimentation on Indigenous CHWs' Motivation

Health and Nutrition

Local food, often known as "field to plate," plays a vital role in connecting Indigenous CHWs to rural health centers. Free from preservatives, pesticides, additives, and flavorings, this food comes straight from the field, offering freshness and abundance, which enhances both its quality and appeal.

> *Whenever people call me to see patients or visit their house, they offer me fresh produce from their farm and sometimes even "desi" (country) chicken for free. Where can you get such nutritious and healthy food in cities?* [Nurse BY, 4-month interview]

Community Engagement

A unique characteristic of Indigenous communities is their emphasis on communitarian living, characterized by strong bonds of sharing and caring for one another [31,32]. Farming serves as both a livelihood and a means of fostering community engagement and identity. Their connection to the land, local markets, and cultural festivals centered around regional cuisine strengthens their sense of belonging and deepens social ties within the community.

> *I visit the villages whenever I have time. During these visits, many people gather to sit and discuss the health and well-being of the community, and we motivate the children. On holidays and Sundays, I often take*

the village youth to the rivers for fishing. [Doctor BA, 4-month interview]

Work-Life Balance

The concern among these CHWs is their inability to manage their domestic chores, as distance limits regular visits to the family and family affairs. The opportunity to serve in their home town facilitates work-life balance and positively impacts their physical and mental health, reducing stress, increasing job satisfaction, and enhancing productivity [33].

Cultural Connection

Food habits often represent a deep cultural bond and sense of belonging [34]. It makes them feel a strong connection to their heritage and traditions through the food they grew up with, making it more appealing to remain in their hometown.

> *We gather together and prepare meals for every celebration in common for all young and old.* [Accredited social health activist PK, 4-month interview]

Retention Trends Among Indigenous CHWs

Recognition

In rural areas, doctors often receive deep respect and appreciation from the rural community. This sense of being valued and recognized enforced emotional fulfillment, encouraging CHWs to continue serving in these regions.

> *I feel like a celebrity, as wherever I go—whether it's the market, the community, or my workplace—people honour and respect me immensely.* [Doctor DM, 4-month interview]

Career Intentions

The state government implemented various strategies to encourage medical students to serve in rural areas, including career growth incentives such as district quotas for entrance into Bachelor of Medicine, Bachelor of Surgery programs; specialized training programs (eg, barefoot doctor training) for rural service; a 3-month community medicine internship in rural settings; government-sponsored quotas for postgraduate, diploma, and degree course selections; as well as the introduction of the Diplomate of National Board program with training conducted in district hospitals [35]. As a result, professional development opportunities, a supportive work environment, community integration, and work-life balance were factors that encouraged CHWs to choose rural areas [36].

> *Once I complete the rural posting then there is an opportunity for further professional growth and other career intentions.* [Doctor SM, 4-month interview]

Promote Local Food and Lifestyle

Access to local food and a lifestyle that aligns with their cultural values and traditions contribute to higher retention rates [37]. The availability of fresh, familiar foods and a slower pace of life compared to urban centers created a more appealing working environment for Indigenous CHWs.

> *When I eat food outside of my region, I face digestion problems. It may be because I am not used to spices and tastemakers. Our tribal food is simple and organic resulting in better health outcomes. Therefore, I prefer to be in rural areas.* [Nurse PK, 6-month interview]

Role of Cultural Beliefs and Practices

The study of sociocultural and economic factors that affected food consumption patterns in Arab countries demonstrates that the cultural beliefs and practices related to food significantly shaped dietary habits and food choices among rural communities [38]. However, in this study, CHWs reported that the ancient practices have a great impact and were driven by a need for local cuisine [39].

Correlations Between Nutritional Support and Job Satisfaction

Better Health and Productivity

Access to nutritional support ensures that health care workers in rural areas stay physically fit and energized, which enhances their job performance [40]. Knowing that their health and well-being will be supported through nutritious, locally sourced food can make rural postings more attractive.

> *I have observed that rural people generally don't suffer from chronic diseases, but rather face issues like accidents, sunburn, sunstroke, or water-borne diseases. We are fortunate to have access to nutritious and healthy food.* [Nurse SH, 4-month interview]

Incentives of Fresh, Organic, and Local Food

Rural areas offer access to fresh, organic, and culturally significant local food. The availability of healthy, farm-to-table meals can serve as a strong motivator for health care workers, making rural postings more appealing due to the unique lifestyle benefits they offer.

> *They don't pay me that time for the treatment I provide when I visit or am called to see patients. They often can't afford to pay, but they give me fresh vegetables, pulses, or fruits that they harvest on the spot. Where else, in urban areas, can you find such genuine incentives and fresh produce?* [Nurse RJ, 6-month interview]

Low Cost of Living

In rural areas, access to fresh, local food can be more affordable than in urban settings. The prospect of spending less on quality food while still enjoying a nutritious diet can make rural postings more financially appealing.

> *I go to a market with 1000 INR [US $15] and buy groceries for the next two to three weeks. Everything is so cheap and fresh in the village markets. Do you think the same in the cities?* [Lab technician AG, 4-month interview]

Ethnicity

The findings demonstrated that ethnicity substantially impacted the food habits of a person owing to traditions, social norms,

migration, and acculturation, which is evident within and outside India [41]. When one travels outside of their home country or region, this becomes quite apparent.

Discussion

Principal Findings

The study findings underscore the positive impact that rural alimentation plays in enhancing the contentment of CHWs and highlight the complex interplay between the rural work environment and the factors that drive their motivation [42]. The results indicate that CHWs with access to nutritious food experienced higher motivation and retention rates [43]. The objective of the study also aligns with previous research showing that psychological factors related to adopting a healthy diet can significantly boost life satisfaction and job motivation. In this study, CHWs expressed satisfaction and a sense of contentment with the availability and quality of food in rural areas. This is similar to the study conducted in Tanzania, which showed that access to nutritious food made CHWs more likely to remain in their positions for extended periods [44].

Role of Nutrition in Enhancing Job Satisfaction

Previous studies have determined that a healthy diet helps protect against many chronic diseases, reducing the risk of developing such conditions [45,46]. The availability of locally sourced, nourishing food enhances rural health care workers' motivation and urges providers and administrators to promote a local and healthy diet, which is a relatively simple and cost-effective strategy to improve CHW motivation and retention [47]. The impact of organic food remains to be determined; it helps reduce food safety risks such as pesticide residue and excessive additives [48].

While there is a strong correlation between nutrition and job satisfaction, few studies, especially in health-related fields, have explored this link. The job satisfaction and food habits of CHWs are largely influenced by their socioeconomic conditions and social and cultural practices. For Indigenous CHWs, local food products play a crucial role in maintaining their health and job satisfaction, which significantly impact their retention [49]. A balanced diet contributes to sustained energy and reduces feelings of fatigue and burnout, allowing workers to perform effectively, which enhances their satisfaction with their jobs. A study on nutra-ergonomics explores the relationship between workers, their work environment, and job satisfaction in connection to their nutritional status. It highlights nutrition as a key component of a safe and productive workplace, influencing physical and mental health, and contributing to long-term retention in their current roles [50].

Cultural and Community Ties

Indigenous peoples typically share a deep ancestral connection to their lands and natural resources. They possess distinct cultures, languages, beliefs, and knowledge systems and maintain strong bonds with their land, properties, and territories. Their unique heritage and traditions are central to their identity and way of life. Culturally and politically, they will find themselves out of place from the rest of society [48].

Impact of Nutritional Support

Nutrition contributes to many indicators of well-being, including maternal health, birth weight, child development, and oral health, and is an important determinant of chronic disease, which reduces life expectancy [51]. Inadequate nutritional intake is a major factor contributing to the burden of disease, and when individuals develop chronic conditions as a result, it often leads to significant out-of-pocket expenses for treatment [52,53].

Government Policy

To attract and retain health workers in rural areas, both the state and central governments have implemented several monetary and nonmonetary benefits:

- Monetary incentives: (1) Hard area allowances and provision of residential facilities; (2) flexible salary schemes, such as the "You Quote, We Pay" strategy, ensuring competitive compensation; and (3) performance-based increments of up to 50% [35,54,55]
- Nonmonetary benefits: (1) Professional development opportunities for doctors, nurses, and allied health workers, including upskilling programs; (2) educational incentives, such as additional National Eligibility cum Entrance Test (Postgraduate) marks—10% for each year of service in remote or difficult areas, up to a maximum of 30%; (3) special honorariums to encourage rural practice among specialists; and (4) reservation of 50% of medical diploma seats for in-service state government doctors who have served in remote or challenging areas

These policies address financial and professional needs, making rural health care roles more attractive and sustainable [35,54]

Implications of the Study

The study revealed several significant implications for the retention and motivation of CHWs in rural settings. It underscored that CHWs with access to nutritious and diverse local food products demonstrated higher motivation and retention rates.

First, enhancing the nutrition of CHWs leads to improved health outcomes within the communities they serve. Given their pivotal role in delivering primary health care services in resource-limited rural areas, ensuring the health and motivation of CHWs directly correlated with the quality of care provided to their communities. Second, addressing the nutritional requirements of CHWs assisted in mitigating the challenge of high turnover rates prevalent in rural areas. CHWs often encounter numerous obstacles that contribute to burnout and turnover, such as long working hours, inadequate remuneration, and inadequate support. Third, the study underscored the significance of tackling social determinants of health, including access to nutritious food, to enhance health care outcomes in underserved communities. By addressing these determinants, health disparities can be reduced, thereby fostering overall community health improvement.

Limitations of the Study

The study was conducted in a specific geographic area and focused on a particular group of CHWs. The study's lack of

robust statistical representation may affect the reliability and generalizability of the results.

Conclusion

The research investigated the relationship between rural alimentation and the motivation of Indigenous CHWs in Jharkhand, India. The findings demonstrated that the retention rates of Indigenous health care workers are positively influenced by their local cuisines and nutrition. Moreover, CHWs with access to organic and locally sourced food exhibited superior retention rates compared to Indigenous CHWs deployed in urban areas. This study also indicated that individuals often exhibit loyalty to their culinary preferences and dietary habits, which drives them to opt for local assignments. Consequently, rural sustenance plays a pivotal role in CHW retention, thereby enhancing the health outcomes of rural residents. In essence, the study underscored the significance of addressing the local diet requirements of CHWs to bolster their motivation and retention rates, consequently elevating the standard of health care services in rural settings. The implications drawn from the study hold crucial insights for policy makers and health care practitioners operating in similar contexts, offering valuable strategies for enhancing the retention and motivation of CHWs in rural areas.

Acknowledgments

The authors extend their heartfelt gratitude to Paul Lelen Hoakip, research scholar, for his invaluable insights and thoughtful contributions, which has greatly enhanced the quality of this paper. Special thanks also to Rev. Dr Ayres Fernandez and Rev. John Thekekara, PhD, for their support and guidance throughout the development of this work. Their expertise and encouragement were instrumental in refining the manuscript and ensuring its academic rigor.

Multimedia Appendix 1
Interview guidelines and questionnaire.
[DOCX File, 19 KB - xmed_v6i1e48346_app1.docx]

References

1. Bitton A, Ratcliffe HL, Veillard JH, et al. Primary health care as a foundation for strengthening health systems in low- and middle-income countries. J Gen Intern Med 2017 May;32(5):566-571. [doi: 10.1007/s11606-016-3898-5] [Medline: 27943038]
2. van Iseghem T, Jacobs I, Vanden Bossche D, et al. The role of community health workers in primary healthcare in the WHO-EU region: a scoping review. Int J Equity Health 2023 Jul 20;22(1):134. [doi: 10.1186/s12939-023-01944-0] [Medline: 37474937]
3. di Renzo L, Gualtieri P, Pivari F, et al. Eating habits and lifestyle changes during COVID-19 lockdown: an Italian survey. J Transl Med 2020 Jun 8;18(1):229. [doi: 10.1186/s12967-020-02399-5] [Medline: 32513197]
4. Aguilera JM. The concept of alimentation and transdisciplinary research. J Sci Food Agric 2021 Mar 30;101(5):1727-1731. [doi: 10.1002/jsfa.10823] [Medline: 32949020]
5. Warra AA, Prasad MNV. Chapter 16 - African perspective of chemical usage in agriculture and horticulture—their impact on human health and environment. In: Prasad MNV, editor. Agrochemicals Detection, Treatment and Remediation: Pesticides and Chemical Fertilizers: Butterworth-Heinemann; 2020:401-436. [doi: 10.1016/B978-0-08-103017-2.00016-7]
6. de Marco M, Thorburn S, Kue J. In a country as affluent as America, people should be eating: experiences with and perceptions of food insecurity among rural and urban Oregonians. Qual Health Res 2009 Jul;19(7):1010-1024. [doi: 10.1177/1049732309338868] [Medline: 19556404]
7. Asfaw A, Simane B, Hassen A, Bantider A. Variability and time series trend analysis of rainfall and temperature in northcentral Ethiopia: a case study in Woleka sub-basin. Weather Clim Extrem 2018 Mar;19:29-41. [doi: 10.1016/j.wace.2017.12.002]
8. Palmer MA, Menninger HL, Bernhardt E. River restoration, habitat heterogeneity and biodiversity: a failure of theory or practice? Freshw Biol 2010 Jan 15;55(s1):205-222. [doi: 10.1111/j.1365-2427.2009.02372.x]
9. National Health Mission. 13th common review mission 2019. National Health Systems Resource Centre. 2019 URL: https://nhsrcindia.org/sites/default/files/2021-04/13th_common_review_mission-Report_2019_Revise.pdf
10. Government of India. Census of India 2011. 2011. URL: https://censusindia.gov.in/nada/index.php/catalog/1366/download/4478/Pesreport.pdf [accessed 2025-01-15]
11. Rural health statistics 2018-19. Ministry of Health and Family Welfare. 2019. URL: https://mohfw.gov.in/?q=reports/rural-health-statistics-2018-19 [accessed 2025-01-15]
12. Kok MC, Kane SS, Tulloch O, et al. How does context influence performance of community health workers in low- and middle-income countries? evidence from the literature. Health Res Policy Sys 2015 Mar 7;13:13. [doi: 10.1186/s12961-015-0001-3] [Medline: 25890229]
13. Jharkhand urban population. Population Census Data (india). 2011. URL: https://www.census2011.co.in/census/state/jharkhand.

html#:~:text=As%20per%20details%20from%20Census,are%2016%2C930%2C315%20and%2016%2C057%2C819%20respectively [accessed 2025-01-17]
14. Fossey E, Harvey C, McDermott F, Davidson L. Understanding and evaluating qualitative research. Aust N Z J Psychiatry 2002 Dec;36(6):717-732. [doi: 10.1046/j.1440-1614.2002.01100.x] [Medline: 12406114]
15. Creswell JW, Creswell JD. Research Design: Qualitative, Quantitative and Mixed Methods Approaches: Sage Publications, Inc; 2009. URL: https://collegepublishing.sagepub.com/products/research-design-6-270550 [accessed 2025-01-17]
16. Sutton J, Austin Z. Qualitative research: data collection, analysis, and management. Can J Hosp Pharm 2015;68(3):226-231. [doi: 10.4212/cjhp.v68i3.1456] [Medline: 26157184]
17. de Leeuw JA, Woltjer H, Kool RB. Identification of factors influencing the adoption of health information technology by nurses who are digitally lagging: in-depth interview study. J Med Internet Res 2020 Aug 14;22(8):e15630. [doi: 10.2196/15630] [Medline: 32663142]
18. Bebbington P, Wilkins S, Sham P, et al. Life events before psychotic episodes: do clinical and social variables affect the relationship? Soc Psychiatry Psychiatr Epidemiol 1996 Jun;31(3-4):122-128. [doi: 10.1007/BF00785758] [Medline: 8766457]
19. Ghai S, Dutta M, Garg A. Perceived level of stress, stressors and coping behaviours in nursing students. Indian J Posit Psychol 2014;5(1):60-65 [FREE Full text]
20. Jacobs K. Discourse analysis. In: Baum S, editor. Methods in Urban Analysis: Springer; 2021:151-172. [doi: 10.1007/978-981-16-1677-8_9]
21. Castleberry A, Nolen A. Thematic analysis of qualitative research data: is it as easy as it sounds? Curr Pharm Teach Learn 2018 Jun;10(6):807-815. [doi: 10.1016/j.cptl.2018.03.019] [Medline: 30025784]
22. Tirandaz Z, Akbarizadeh G, Kaabi H. PolSAR image segmentation based on feature extraction and data compression using weighted neighborhood filter bank and hidden Markov random field-expectation maximization. Measurement (Lond) 2020 Mar;153:107432. [doi: 10.1016/j.measurement.2019.107432]
23. Bradley EH, Curry LA, Devers KJ. Qualitative data analysis for health services research: developing taxonomy, themes, and theory. Health Serv Res 2007 Aug;42(4):1758-1772. [doi: 10.1111/j.1475-6773.2006.00684.x] [Medline: 17286625]
24. Thomas J, Harden A. Methods for the thematic synthesis of qualitative research in systematic reviews. BMC Med Res Methodol 2008 Jul 10;8:45. [doi: 10.1186/1471-2288-8-45] [Medline: 18616818]
25. Baškarada S. Qualitative case study guidelines. The Qualitative Report 2014;19(40):1-18. [doi: 10.46743/2160-3715/2014.1008]
26. Mbindyo P, Gilson L, Blaauw D, English M. Contextual influences on health worker motivation in district hospitals in Kenya. Implementation Sci 2009 Jul 23;4:43. [doi: 10.1186/1748-5908-4-43] [Medline: 19627590]
27. Coventry PA, Fisher L, Kenning C, Bee P, Bower P. Capacity, responsibility, and motivation: a critical qualitative evaluation of patient and practitioner views about barriers to self-management in people with multimorbidity. BMC Health Serv Res 2014 Oct 31;14(1):536. [doi: 10.1186/s12913-014-0536-y] [Medline: 25367263]
28. Cavaye ALM. Case study research: a multi‐faceted research approach for IS. Information Systems Journal 1996;6(3):227-242. [doi: 10.1111/j.1365-2575.1996.tb00015.x]
29. Peterson BL. Thematic analysis/interpretive thematic analysis. In: The International Encyclopedia of Communication Research Methods: Wiley; 2017:1-9. [doi: 10.1002/9781118901731.iecrm0249]
30. Willan J. Doing early childhood research. In: Early Childhood Studies: A Multidisciplinary Approach: Bloomsbury Publishing; 2017:353-371 URL: https://www.bloomsbury.com/ca/early-childhood-studies-9781137274021/ [accessed 2025-01-23] [doi: 10.1057/978-1-137-27402-1_18]
31. Coates J. In: Gray M, editor. Indigenous Social Work around the World: Towards Culturally Relevant Education and Practice: Routledge; 2016:1-5. [doi: 10.4324/9781315588360]
32. Smith-Morris C. Indigenous Communalism: Belonging, Healthy Communities, and Decolonizing the Collective: Rutgers University Press; 2020:25. [doi: 10.2307/j.ctvscxrb6.5]
33. Akintola O, Chikoko G. Factors influencing motivation and job satisfaction among supervisors of community health workers in marginalized communities in South Africa. Hum Resour Health 2016 Sep 6;14(1):54. [doi: 10.1186/s12960-016-0151-6] [Medline: 27601052]
34. Weber EU. Climate change demands behavioral change: what are the challenges? Social Research: An International Quarterly 2015;82(3):561-580. [doi: 10.1353/sor.2015.0050]
35. Ghosh K. Why we don't get doctors for rural medical service in India? Natl Med J India 2018;31(1):44-46. [doi: 10.4103/0970-258X.243416] [Medline: 30348926]
36. Agyapong VIO, Osei A, Farren CK, McAuliffe E. Factors influencing the career choice and retention of community mental health workers in Ghana. Hum Resour Health 2015 Jul 9;13(1):56. [doi: 10.1186/s12960-015-0050-2] [Medline: 26156234]
37. Schrank Z, Running K. Individualist and collectivist consumer motivations in local organic food markets. J Cons Cult 2018 Feb;18(1):184-201. [doi: 10.1177/1469540516659127]
38. Musaiger AO. Socio-cultural and economic factors affecting food consumption patterns in the Arab countries. J R Soc Health 1993 Apr;113(2):68-74. [doi: 10.1177/146642409311300205] [Medline: 8478894]

39. Chakona G. Social circumstances and cultural beliefs influence maternal nutrition, breastfeeding and child feeding practices in South Africa. Nutr J 2020 May 20;19(1):47. [doi: 10.1186/s12937-020-00566-4] [Medline: 32434557]
40. Karaferis D, Aletras V, Raikou M, Niakas D. Factors influencing motivation and work engagement of healthcare professionals. Mater Sociomed 2022 Sep;34(3):216-224. [doi: 10.5455/msm.2022.34.216-224] [Medline: 36310751]
41. Verbeke W, Poquiviqui López G. Ethnic food attitudes and behaviour among Belgians and Hispanics living in Belgium. British Food Journal 2005 Dec 1;107(11):823-840. [doi: 10.1108/00070700510629779]
42. Wronska MD, Coffey M, Robins A. Determinants of nutrition practice and food choice in UK construction workers. Health Promot Int 2022 Oct 1;37(5):daac129. [doi: 10.1093/heapro/daac129] [Medline: 36166265]
43. Ozano K, Simkhada P, Thann K, Khatri R. Improving local health through community health workers in Cambodia: challenges and solutions. Hum Resour Health 2018 Jan 6;16(1):2. [doi: 10.1186/s12960-017-0262-8] [Medline: 29304869]
44. Greenspan JA, McMahon SA, Chebet JJ, Mpunga M, Urassa DP, Winch PJ. Sources of community health worker motivation: a qualitative study in Morogoro Region, Tanzania. Hum Resour Health 2013 Oct 10;11:52. [doi: 10.1186/1478-4491-11-52] [Medline: 24112292]
45. Steinman L, Heang H, van Pelt M, et al. Facilitators and barriers to chronic disease self-management and mobile health interventions for people living with diabetes and hypertension in Cambodia: qualitative study. JMIR Mhealth Uhealth 2020 Apr 24;8(4):e13536. [doi: 10.2196/13536] [Medline: 32329737]
46. Cena H, Calder PC. Defining a healthy diet: evidence for the role of contemporary dietary patterns in health and disease. Nutrients 2020 Jan 27;12(2):334. [doi: 10.3390/nu12020334] [Medline: 32012681]
47. Mozaffarian D, Angell SY, Lang T, Rivera JA. Role of government policy in nutrition-barriers to and opportunities for healthier eating. BMJ 2018 Jun 13;361:k2426. [doi: 10.1136/bmj.k2426] [Medline: 29898890]
48. Chai D, Meng T, Zhang D. Influence of food safety concerns and satisfaction with government regulation on organic food consumption of Chinese urban residents. Foods 2022 Sep 22;11(19):2965. [doi: 10.3390/foods11192965] [Medline: 36230045]
49. Kendirkiran G, Batur B. The relationship between eating behavior and job satisfaction of academic staff. Int J Caring Sci 2022 Aug;15(2):825-836 [FREE Full text]
50. Shearer J, Graham TE, Skinner TL. Nutra-ergonomics: influence of nutrition on physical employment standards and the health of workers. Appl Physiol Nutr Metab 2016 Jun;41(6 Suppl 2):S165-S174. [doi: 10.1139/apnm-2015-0531] [Medline: 27277565]
51. Alfred T, Corntassel J. Being Indigenous: resurgences against contemporary colonialism. Government and Opposition 2005;40(4):597-614. [doi: 10.1111/j.1477-7053.2005.00166.x]
52. Browne J, Hayes R, Gleeson D. Aboriginal health policy: is nutrition the 'gap' in 'Closing the Gap'? Aust N Z J Public Health 2014 Aug;38(4):362-369. [doi: 10.1111/1753-6405.12223] [Medline: 25091077]
53. Schembri L, Curran J, Collins L, et al. The effect of nutrition education on nutrition‐related health outcomes of Aboriginal and Torres Strait Islander people: a systematic review. Aust N Z J Public Health 2016 Apr;40(Suppl 1):S42-S47. [doi: 10.1111/1753-6405.12392] [Medline: 26123037]
54. Press Information Bureau, Ministry of Information and Broadcasting, Government of India. Revolutionizing healthcare: digital innovations in India's health sector. Press Information Bureau. 2024 Jan 15. URL: https://static.pib.gov.in/WriteReadData/specificdocs/documents/2024/jan/doc2024115298601.pdf [accessed 2025-01-15]
55. Silvestri DM, Blevins M, Afzal A, et al. Medical and nursing students' intentions to work abroad or in rural areas: an eight-country cross-sectional survey in Asia and Africa. Ann Glob Health 2015 Mar 12;81(1):52. [doi: 10.1016/j.aogh.2015.02.627]

Abbreviations

CHC: Community Health Centre
CHW: community health worker

Edited by A Schwartz; submitted 20.04.23; peer-reviewed by SK Thalari; revised version received 28.11.24; accepted 13.12.24; published 23.01.25.

Please cite as:
Kerketta A, A N R
The Impact of Rural Alimentation on the Motivation and Retention of Indigenous Community Health Workers in India: A Qualitative Study
JMIRx Med 2025;6:e48346
URL: https://xmed.jmir.org/2025/1/e48346
doi:10.2196/48346

Applications of Indocyanine Green in Breast Cancer for Sentinel Lymph Node Mapping: Protocol for a Scoping Review

Feryal Kurdi*, MD; Yahya Kurdi*, MD; Igor Vladimirovich Reshetov, MD, PhD

Department of Oncology, Radiotherapy and Plastic and Reconstructive Surgery, Sechenov University, Bolshaya Pirogovskaya, 6c1, Moscow, Russian Federation

*these authors contributed equally

Corresponding Author:
Feryal Kurdi, MD
Department of Oncology, Radiotherapy and Plastic and Reconstructive Surgery, Sechenov University, Bolshaya Pirogovskaya, 6c1, Moscow, Russian Federation

Related Articles:

Companion article: https://www.medrxiv.org/content/10.1101/2024.07.30.24311256v1

Companion article: https://med.jmirx.org/2025/1/e69705

Companion article: https://med.jmirx.org/2025/1/e68769

Abstract

Introduction: Breast cancer is the leading cause of morbidity and mortality worldwide. Accurate sentinel lymph node (SLN) mapping is crucial for staging and treatment planning in early-stage breast cancer. Indocyanine green (ICG) has emerged as a promising agent for fluorescence imaging in SLN mapping. However, comprehensive assessment of its clinical utility, including accuracy and adverse effects, remains limited. This scoping review aims to consolidate evidence on the use of ICG in breast cancer SLN mapping.

Objective: The objective of this scoping review is to evaluate the current literature on the use of ICG in SLN mapping for patients with breast cancer. This review aims to assess the accuracy, efficacy, and safety of ICG in this context and to identify gaps in the existing research. The outcomes will contribute to the development of further research as part of a PhD project.

Methods: Five electronic databases will be searched (PubMed, Embase, MEDLINE, Web of Science, and Scopus) using search strategies developed in consultation with an academic supervisor. The search strategy is set to human studies published in English within the last 11 years. All retrieved citations will be imported to Zotero and then uploaded to Covidence for the screening of titles, abstracts, and full text according to prespecified inclusion criteria. Patients with early-stage breast cancer (T1 and T2), selected T3 cases where the SLN biopsy is accurate, and those with clinically node-negative breast cancer will be included. The intervention criterion includes studies using ICG for SLN mapping and studies on the assessment of fluorescence imaging cameras. Citations meeting the inclusion criteria for full-text review will have their data extracted by 2 independent reviewers, with disagreements resolved by discussion. A data extraction tool will be developed to capture full details about the participants, concept, and context, and findings relevant to the scoping review will be summarized.

Results: The preliminary search began in December 2023. As of September 2024, papers have been screened and data are currently being extracted. Out of the 2130 references initially imported, 126 studies met the inclusion criteria after screening. The scoping review is expected to be published in January 2025.

Conclusions: Although ICG technology has been used for SLN mapping in patients with breast cancer, initial searches in 2022 revealed limited data on this technique's feasibility, safety, and effectiveness. At that time, preliminary search of Scopus, MEDLINE, Embase, and PubMed identified no current or forthcoming systematic reviews or scoping reviews on the topic. However, recent searches indicate a substantial increase in research and reviews, reflecting a growing interest and evidence in this area.

(JMIRx Med 2025;6:e66213) doi:10.2196/66213

KEYWORDS

indocyanine green; ICG; sentinel lymph node; breast cancer; fluorescence; axillary lymph node mapping; NIR; surgical planning; near-infrared

Introduction

Sentinel lymph node (SLN) biopsy plays a crucial role in staging and prognosis in breast cancer management. The SLN is the initial lymph node to which breast cancer cells are likely to metastasize, and the presence of cancer cells in the SLN indicates a higher likelihood of further metastasis to other lymph nodes and distant organs [1].

SLN biopsy involves injecting a tracer substance into the breast, which then migrates to the SLN. The SLN is then identified, excised, and examined for cancer cells. If the SLN is free of cancer cells, it suggests that the cancer has not spread to other lymph nodes, eliminating the need for additional lymph node dissection. Conversely, if the SLN contains metastases, further dissection is typically required [2].

Over the past 2 decades, SLN biopsy using blue dye and radiotracers has been established as the diagnostic standard of care for patients with early-stage breast cancer who have clinically negative lymph nodes [3,4].

However, these methods come with certain drawbacks, including the potential for allergic reactions to the blue dye and the necessity of nuclear medicine facilities for radiotracer injection and detection. In a cohort undergoing blue dye and radiotracer injection procedures, a small number of adverse reactions, such as skin tattooing and anaphylaxis, were reported [5].

In recent years, near-infrared (NIR) fluorescence imaging using indocyanine green (ICG) has emerged as an alternative approach for SLN mapping in patients with breast cancer. ICG, a fluorescent dye, is injected into the breast, which then migrates to the SLNs. A NIR camera detects the fluorescence emitted by ICG, enabling the surgeon to identify and excise the SLNs [6,7].

This technology offers several advantages over traditional methods, including enhanced visualization of SLNs, a lower risk of allergic reactions, and the elimination of the need for nuclear medicine facilities. Furthermore, ICG has an excellent safety profile [8-11].

The importance of this topic stems from the potential of ICG technology to enhance the accuracy and safety of SLN mapping in patients with breast cancer. Precise identification and removal of the SLN are crucial for accurate staging and prognosis. Inaccurate SLN identification can lead to unnecessary lymph node dissection, resulting in complications such as lymphedema and impaired arm function. Sampling a larger number of SLNs may increase the risk of upper limb lymphedema, sensory deficits, and reduced shoulder function.

Landmark trials have shown a significant difference in morbidity rates when comparing SLN biopsy to axillary dissection, with rates of 25% and 70%, respectively [3,12]. Recent studies have reported excising, on average, 2 nodes per patient, likely due to advancements in NIR technology and ICG fluorescence protocols [13-17]. Nevertheless, further research is essential to assess the long-term outcomes and cost-effectiveness of ICG technology compared to traditional methods.

Methods

Overview

The proposed scoping review will be guided by the JBI methodology for scoping reviews [18]. The search strategy aims to locate both published and unpublished articles. An initial limited search of PubMed, Embase, MEDLINE, Web of Science, and Scopus was undertaken to identify relevant articles on the use of ICG for SLN mapping in breast cancer. In consultation with an academic supervisor, the keywords in the titles and abstracts of relevant articles, as well as the index terms used to describe these articles, were used to develop a comprehensive search strategy for PubMed, Embase, MEDLINE, Web of Science, and Scopus (see Multimedia Appendix 1). This strategy, including all identified keywords and index terms, will be adapted for each included database. The articles sourced from all included sources of evidence will be exported into Zotero (Corporation for Digital Scholarship).

Only articles published in English will be included due to the language proficiency of the reviewers. Articles published since January 1, 2014, will be included to ensure relevance, aligning with the project's consideration of recent data and the ongoing advancements in SLN mapping techniques using ICG.

JBI Methodology for Scoping Reviews

The outcomes of the scoping review will inform and frame three subsequent pieces of work planned as part of a PhD project:

1. Prospective cohort study on the long-term outcomes of ICG in SLN mapping
2. Systematic review and meta-analysis of ICG for SLN mapping in breast cancer
3. Development of standardized clinical guidelines and protocols for the use of ICG in SLN mapping in patients with breast cancer

The Participants-Concept-Context framework for this scoping review defines (1) the participants as patients with early-stage breast cancer, (2) the concept as the use of ICG for SLN mapping in patients with breast cancer, and (3) the context as SLN mapping that is performed as part of breast cancer staging and treatment planning.

Review Questions

The review questions are as follows:

1. What do we know about the evaluation and integration of emergent evidence on the use of ICG for SLN mapping in patients with breast cancer into clinical practice and decision-making?
2. To what extent is emergent evidence on the feasibility, safety, and effectiveness of ICG for SLN mapping integrated into clinical guidelines and decision-making processes?
3. How is emergent evidence on the use of ICG for SLN mapping evaluated and incorporated into clinical guidelines and decision-making processes?

For the purposes of this scoping review, emergent evidence refers to new research findings on ICG for SLN mapping that

have emerged after market launch and have not yet been fully integrated into clinical guidelines and practice.

Eligibility Criteria

The eligibility criteria are as follows. Participants will include patients with early-stage breast cancer (T1 and T2) and selected T3 cases where SLN biopsy has been shown to be accurate. Concept will include the use of ICG for SLN mapping in patients with breast cancer, as well as the assessment of imaging techniques and devices used in conjunction with ICG for SLN mapping. Context will include clinical settings where SLN mapping is performed as part of breast cancer staging and treatment planning.

This scoping review will consider both experimental and quasi-experimental study designs, including controlled before-and-after studies and controlled interrupted time-series studies. In addition, analytical observational studies including prospective and retrospective cohort studies, case-control studies, and analytical cross-sectional studies will be considered for inclusion. This review will also consider descriptive observational study designs such as descriptive cross-sectional studies for inclusion. Qualitative studies that focus on qualitative data will be considered for inclusion.

Following the search, all identified articles will be exported into Zotero. Then, the remaining articles will be uploaded into Covidence (Veritas Health Innovations Ltd). Titles and abstracts will then be screened by the lead author against the inclusion criteria for the scoping review. Potentially relevant articles will be retrieved in full and included in Covidence. The full text of these articles will be assessed in detail against the inclusion criteria by 2 independent reviewers. Reasons for the exclusion of sources of evidence at the full-text stage that do not meet the inclusion criteria will be recorded and reported in the scoping review. Any disagreements that arise between the reviewers at each stage of the selection process will be resolved through discussion or with an additional reviewer. The results of the search and the study inclusion process will be reported in full in the final scoping review and presented in a PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews) checklist, as extracted from Covidence (see Multimedia Appendix 2) [19].

Data will be extracted from all articles included in the scoping review by 2 independent reviewers, using a data extraction tool developed by the lead reviewer and piloted with about 15 articles to refine and improve it. The data extracted will include specific details about the participants, concept, context, study methods, and key findings relevant to the scoping review questions and will be imported into either Covidence or Microsoft Excel.

A draft extraction form is provided (see Multimedia Appendix 3). The draft data extraction tool will be modified and revised as necessary during the process of extracting data from each included article. Modifications will be detailed in the scoping review. Any disagreements that arise between the reviewers will be resolved through discussion or with an additional reviewer. If appropriate, authors of articles will be contacted to request missing or additional data, where required.

The evidence presented will directly respond to the scoping review's objective and questions. The data will be presented graphically or in diagrammatic or tabular form. A narrative summary will accompany the tabulated and/or charted results and will describe how the results relate to the scoping review's objective and questions.

Results

The preliminary search began in December 2023. As of September 2024, papers have been screened and data are currently being extracted. Out of the 2130 references initially imported, 126 studies met the inclusion criteria after screening (see Multimedia Appendix 4). The scoping review is anticipated to be published in January 2025.

Discussion

The significance of SLN mapping using ICG technology in breast cancer lies in its potential to enhance accuracy and safety, reduce complications, and improve patient outcomes [20]. Although ICG technology has been used for SLN mapping in patients with breast cancer, initial searches in 2022 revealed limited data on the feasibility, safety, and effectiveness of this technique. At that time, a preliminary search of Scopus, MEDLINE, Embase, and PubMed identified no current or forthcoming systematic reviews or scoping reviews on the topic. However, recent searches indicate a substantial increase in research and reviews, reflecting a growing interest and evidence in this area. Further studies are necessary to assess the long-term efficacy and cost-effectiveness of this technique and to identify the patient populations most likely to benefit.

The objective of this scoping review is to assess the extent of the literature on SLN mapping using ICG technology around the evaluation and integration of emergent evidence for benefits and harms; explore its feasibility, safety, and effectiveness in a larger cohort of patients with breast cancer; and provide guidance for clinical decision-making.

This scoping review could also identify specific patient populations, such as those with higher BMIs, who may benefit most from ICG technology. Additionally, patients who have undergone neoadjuvant therapy could be particularly advantageous candidates.

Factors such as the type of NIR cameras used, the learning curve for surgeons to become proficient with ICG for SLN detection, the availability of ICG and radioisotopes, the presence of nuclear medicine facilities, regional variations in ICG usage, and cost comparisons with the gold standard are also critical considerations in the broader adoption of this technology.

Limitations of this study include a lack of quantitative synthesis (ie, meta-analysis) of the results, which may limit the ability to draw strong conclusions. This scoping review serves as a foundational step toward a more comprehensive systematic review and meta-analysis guiding the clinical decision-making and the integration of ICG into standardized guidelines for SLN mapping in patients with breast cancer.

Acknowledgments

This scoping review is to contribute in part to a Doctor of Philosophy degree.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Search strategy.
[DOCX File, 14 KB - xmed_v6i1e66213_app1.docx]

Multimedia Appendix 2
PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews) checklist.
[DOCX File, 112 KB - xmed_v6i1e66213_app2.docx]

Multimedia Appendix 3
Data extraction instrument.
[DOCX File, 15 KB - xmed_v6i1e66213_app3.docx]

Multimedia Appendix 4
PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) extraction flowchart.
[DOCX File, 79 KB - xmed_v6i1e66213_app4.docx]

References

1. McMasters KM, Tuttle TM, Carlson DJ, et al. Sentinel lymph node biopsy for breast cancer: a suitable alternative to routine axillary dissection in multi-institutional practice when optimal technique is used. J Clin Oncol 2000 Jul;18(13):2560-2566. [doi: 10.1200/JCO.2000.18.13.2560] [Medline: 10893287]
2. Gradishar WJ, Moran MS, Abraham J, et al. NCCN Guidelines® Insights: Breast Cancer, version 4.2023. J Natl Compr Canc Netw 2023 Jun;21(6):594-608. [doi: 10.6004/jnccn.2023.0031] [Medline: 37308117]
3. Mansel RE, Fallowfield L, Kissin M, et al. Randomized multicenter trial of sentinel node biopsy versus standard axillary treatment in operable breast cancer: the ALMANAC Trial. J Natl Cancer Inst 2006 May 3;98(9):599-609. [doi: 10.1093/jnci/djj158] [Medline: 16670385]
4. Donker M, van Tienhoven G, Straver ME, et al. Radiotherapy or surgery of the axilla after a positive sentinel node in breast cancer (EORTC 10981-22023 AMAROS): a randomised, multicentre, open-label, phase 3 non-inferiority trial. Lancet Oncol 2014 Nov;15(12):1303-1310. [doi: 10.1016/S1470-2045(14)70460-7] [Medline: 25439688]
5. Nguyen CL, Zhou M, Easwaralingam N, et al. Novel dual tracer indocyanine green and radioisotope versus gold standard sentinel lymph node biopsy in breast cancer: the GREENORBLUE Trial. Ann Surg Oncol 2023 Oct;30(11):6520-6527. [doi: 10.1245/s10434-023-13824-6] [Medline: 37402976]
6. Polom K, Murawa D, Rho YS, Nowaczyk P, Hünerbein M, Murawa P. Current trends and emerging future of indocyanine green usage in surgery and oncology: a literature review. Cancer 2011 Nov 1;117(21):4812-4822. [doi: 10.1002/cncr.26087] [Medline: 21484779]
7. Liberale G, Vankerckhove S, Bouazza F, et al. Systemic sentinel lymph node detection using fluorescence imaging after indocyanine green intravenous injection in colorectal cancer: protocol for a feasibility study. JMIR Res Protoc 2020 Aug 14;9(8):e17976. [doi: 10.2196/17976] [Medline: 32554370]
8. van der Vorst JR, Schaafsma BE, Hutteman M, et al. Near-infrared fluorescence-guided resection of colorectal liver metastases. Cancer 2013 Sep 15;119(18):3411-3418. [doi: 10.1002/cncr.28203] [Medline: 23794086]
9. Hope-Ross M, Yannuzzi LA, Gragoudas ES, et al. Adverse reactions due to indocyanine green. Ophthalmology 1994 Mar;101(3):529-533. [doi: 10.1016/s0161-6420(94)31303-0] [Medline: 8127574]
10. Griffiths M, Chae MP, Rozen WM. Indocyanine green-based fluorescent angiography in breast reconstruction. Gland Surg 2016 Apr;5(2):133-149. [doi: 10.3978/j.issn.2227-684X.2016.02.01] [Medline: 27047782]
11. Benya R, Quintana J, Brundage B. Adverse reactions to indocyanine green: a case report and a review of the literature. Cathet Cardiovasc Diagn 1989 Aug;17(4):231-233. [doi: 10.1002/ccd.1810170410] [Medline: 2670244]
12. Giuliano AE, Hunt KK, Ballman KV, et al. Axillary dissection vs no axillary dissection in women with invasive breast cancer and sentinel node metastasis: a randomized clinical trial. JAMA 2011 Feb 9;305(6):569-575. [doi: 10.1001/jama.2011.90] [Medline: 21304082]

13. Fleissig A, Fallowfield LJ, Langridge CI, et al. Post-operative arm morbidity and quality of life. results of the ALMANAC randomised trial comparing sentinel node biopsy with standard axillary treatment in the management of patients with early breast cancer. Breast Cancer Res Treat 2006 Feb;95(3):279-293. [doi: 10.1007/s10549-005-9025-7] [Medline: 16163445]
14. Ahmed M, Purushotham AD, Douek M. Novel techniques for sentinel lymph node biopsy in breast cancer: a systematic review. Lancet Oncol 2014 Jul;15(8):e351-e362. [doi: 10.1016/S1470-2045(13)70590-4] [Medline: 24988938]
15. Schaafsma BE, Verbeek FPR, Rietbergen DDD, et al. Clinical trial of combined radio- and fluorescence-guided sentinel lymph node biopsy in breast cancer. Br J Surg 2013 Jul;100(8):1037-1044. [doi: 10.1002/bjs.9159] [Medline: 23696463]
16. Ballardini B, Santoro L, Sangalli C, et al. The indocyanine green method is equivalent to the 99mTc-labeled radiotracer method for identifying the sentinel node in breast cancer: a concordance and validation study. Eur J Surg Oncol 2013 Dec;39(12):1332-1336. [doi: 10.1016/j.ejso.2013.10.004] [Medline: 24184123]
17. Abe H, Yamazaki K, Tokuda A, Ogawa M, Kawasaki M, Kameyama M. A novel approach for sentinel lymph node identification using fluororescence imaging and computed tomography lymphography in early-stage breast cancer patients. J Clin Oncol 2014 May 20;32(15_suppl):e12025-e12025. [doi: 10.1200/jco.2014.32.15_suppl.e12025]
18. Peters MDJ, Marnie C, Tricco AC, et al. Updated methodological guidance for the conduct of scoping reviews. JBI Evid Synth 2020 Oct;18(10):2119-2126. [doi: 10.11124/JBIES-20-00167] [Medline: 33038124]
19. Peters MDJ, Godfrey CM, Khalil H, McInerney P, Parker D, Soares CB. Guidance for conducting systematic scoping reviews. Int J Evid Based Healthc 2015 Sep;13(3):141-146. [doi: 10.1097/XEB.0000000000000050] [Medline: 26134548]
20. Sugie T, Ikeda T, Kawaguchi A, Shimizu A, Toi M. Sentinel lymph node biopsy using indocyanine green fluorescence in early-stage breast cancer: a meta-analysis. Int J Clin Oncol 2017 Feb;22(1):11-17. [doi: 10.1007/s10147-016-1064-z] [Medline: 27864624]

Abbreviations

ICG: indocyanine green
NIR: near-infrared
PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews
SLN: sentinel lymph node

Edited by S Tungjitviboonkun; submitted 06.09.24; peer-reviewed by Anonymous; revised version received 20.10.24; accepted 21.10.24; published 06.01.25.

Please cite as:
Kurdi F, Kurdi Y, Reshetov IV
Applications of Indocyanine Green in Breast Cancer for Sentinel Lymph Node Mapping: Protocol for a Scoping Review
JMIRx Med 2025;6:e66213
URL: https://xmed.jmir.org/2025/1/e66213
doi:10.2196/66213

Cardiotoxicity in Pediatric Cancer Survivorship: Retrospective Cohort Study

Masab Mansoor[1], DBA; Andrew Ibrahim[2], BS

[1]Edward Via College of Osteopathic Medicine, 4408 Bon Aire Drive, Monroe, LA, United States

[2]Texas Tech University Health Sciences Center School of Medicine, Lubbock, TX, United States

Corresponding Author:
Masab Mansoor, DBA
Edward Via College of Osteopathic Medicine, 4408 Bon Aire Drive, Monroe, LA, United States

Related Articles:

Companion article: https://www.medrxiv.org/content/10.1101/2024.08.10.24311795v1

Companion article: https://med.jmirx.org/2025/1/e79521

Companion article: https://med.jmirx.org/2025/1/e79523

Companion article: https://med.jmirx.org/2025/1/e79672

Abstract

Background: Improved survival rates in pediatric cancer have shifted focus to long-term effects of treatment, with cardiovascular complications emerging as a leading cause of morbidity and mortality. Understanding the patterns and predictors of cardiotoxicity is crucial for risk stratification, treatment optimization, and long-term care planning.

Objective: This study investigated the prevalence, incidence, and risk factors of cardiotoxicity in pediatric cancer survivors using data from the Childhood Cancer Survivor Study.

Methods: We conducted a retrospective cohort study of 24,938 five-year survivors of childhood cancer diagnosed between 1970 and 1999. Cardiovascular complications, including cardiomyopathy, coronary artery disease, valvular heart disease, and arrhythmias, were assessed through self-reported questionnaires and medical record review. Cox proportional hazards models were used to evaluate risk factors, and a prediction model was developed using multivariable logistic regression.

Results: The cumulative incidence of any cardiovascular complication by 30 years postdiagnosis was 18.7% (95% CI 17.9% - 19.5%). Significant risk factors included anthracycline exposure (hazard ratio 2.31, 95% CI 2.09 - 2.55 for doses ≥250 mg/m²), chest radiation (hazard ratio 1.84, 95% CI 1.66 - 2.05 for doses ≥20 Gy), older age at diagnosis, male sex, and obesity. A risk prediction model demonstrated good discrimination (C statistic 0.78, 95% CI 0.76 - 0.80). Survivors had a significantly higher risk of cardiovascular complications compared with sibling controls (odds ratio 3.7, 95% CI 3.2 - 4.2).

Conclusions: Childhood cancer survivors face a substantial and persistent risk of cardiovascular complications. The identified risk factors and prediction model can guide personalized follow-up strategies and interventions. These findings underscore the need for lifelong cardiovascular monitoring and care in this population.

(JMIRx Med 2025;6:e65299) doi:10.2196/65299

KEYWORDS

cardiotoxicity; cardiology; cardiovascular; heart; arrhythmias; self-reported questionnaires; oncology; survivors; pediatrics; prevalence; incidence; risk; epidemiology; anthracycline exposure; childhood cancer survivors

Introduction

Background

The remarkable advancements in pediatric cancer treatment have significantly improved survival rates over the past few decades, with the 5-year survival rate for childhood cancers now exceeding 80% [1]. This success has shifted the focus toward understanding and mitigating the long-term effects of cancer treatments on survivors. Among these late effects, cardiovascular complications have emerged as a leading cause of morbidity and mortality in childhood cancer survivors [2].

Cardiotoxicity, a term encompassing a spectrum of cardiovascular adverse effects, can manifest in various forms including cardiomyopathy, coronary artery disease, valvular heart disease, and arrhythmias [3]. Multiple factors such as type of cancer, treatment modalities, and patient-specific characteristics [4] influence the risk of developing these complications.

Anthracyclines, a class of chemotherapeutic agents widely used in pediatric oncology, are particularly associated with cardiotoxicity [5]. While their efficacy in treating various childhood cancers is well-established, the potential for long-term cardiac damage poses a significant challenge in balancing treatment efficacy with long-term health outcomes [6]. Radiation therapy, especially when the heart is within the treatment field, also contributes to increased cardiovascular risk in survivors [7].

The temporal pattern of cardiotoxicity presentation varies, with some effects appearing during or shortly after treatment, while others may not manifest until decades later [8]. This delayed onset presents unique challenges in long-term care and monitoring of childhood cancer survivors.

Understanding the patterns and predictors of cardiotoxicity is crucial for several reasons, including risk stratification (identifying high-risk individuals allows for targeted surveillance and early intervention) [9], treatment optimization (balancing oncological efficacy with cardioprotection in future treatment protocols) [10], long-term care planning (developing evidence-based guidelines for cardiovascular monitoring and management in survivors) [4], and patient education (empowering survivors with knowledge about potential risks and preventive strategies) [11].

The Childhood Cancer Survivor Study (CCSS), a multi-institutional, longitudinal cohort study, provides a robust platform for investigating these long-term health outcomes [12]. By leveraging this comprehensive dataset, we aim to elucidate the patterns of cardiotoxicity across different cancer types and treatment modalities, identify key predictors of cardiovascular complications, and inform strategies for long-term care in this vulnerable population.

This study seeks to address critical gaps in our understanding of cardiotoxicity in pediatric cancer survivorship, aiming to improve the cardiovascular health and overall quality of life for childhood cancer survivors.

Objectives

The primary aim of this study is to comprehensively investigate cardiotoxicity in pediatric cancer survivors using data from the Childhood Cancer Survivor Study (CCSS; Textbox 1).

Textbox 1. Objectives of this study.

- Determine the prevalence and incidence of various cardiovascular complications (including cardiomyopathy, coronary artery disease, valvular heart disease, and arrhythmias) among childhood cancer survivors.
- Analyze the temporal patterns of cardiotoxicity onset in relation to cancer diagnosis and treatment completion.
- Identify and quantify the impact of potential risk factors for cardiotoxicity, including cancer type, treatment modalities (eg, specific chemotherapy agents, cumulative anthracycline dose, and radiation therapy), patient characteristics (eg, age at diagnosis, sex, and genetic predisposition), and lifestyle factors (eg, obesity, physical activity, and smoking status).
- Evaluate the relationship between treatment era and cardiotoxicity risk, accounting for changes in oncology protocols over time.
- Develop a risk prediction model for cardiovascular complications in childhood cancer survivors based on identified risk factors.
- Assess the impact of cardiovascular complications on overall survival and quality of life measures in the survivor cohort.
- Explore potential cardioprotective factors or interventions associated with reduced risk of cardiovascular complications.
- Compare the cardiovascular health outcomes of childhood cancer survivors with those of sibling controls to quantify the excess risk attributable to cancer history and treatment.

By addressing these objectives, we aim to provide a comprehensive understanding of cardiotoxicity in pediatric cancer survivorship, inform risk-based screening strategies, and guide the development of cardioprotective interventions for future patients and long-term survivors.

Methods

Study Population and Data Source

We conducted a retrospective cohort study using data from the CCSS. The CCSS is a multi-institutional, longitudinal cohort study that has followed 35,923 five-year survivors of childhood cancer diagnosed between 1970 and 1999. Eligible participants were those diagnosed with cancer before the age of 21 years, who were treated at 1 of the 31 participating institutions across the United States and Canada [13]. These institutions collectively represent major pediatric oncology centers, providing comprehensive coverage across North America. The CCSS cohort represents one of the largest and most comprehensive resources for studying long-term outcomes in childhood cancer survivors. We included all participants with complete data on cardiovascular outcomes and relevant treatment information.

Outcome Measures

The primary outcomes of interest were cardiovascular complications, including cardiomyopathy, coronary artery disease, valvular heart disease, and arrhythmias. These outcomes were ascertained through self-reported questionnaires. To enhance validity, 27% of all self-reported cardiovascular events (739 of 2743 cases) were confirmed through medical record review by trained abstractors using standardized protocols [14].

The validation procedure showed a 93% confirmation rate for self-reported cardiovascular conditions.

Exposure Variables

We collected data on cancer diagnosis (type and stage), treatment modalities (chemotherapy agents and cumulative doses, radiation therapy [fields and doses], and surgical interventions), patient characteristics (age at diagnosis, sex, race or ethnicity, and family history of cardiovascular disease), and lifestyle factors (BMI, physical activity level, and smoking status).

Data Analysis

Statistical analyses were performed using R (version 4.1.0; R Foundation for Statistical Computing). Descriptive statistics were calculated for all variables. Continuous variables were summarized as means (SD) or medians (IQR), and categorical variables as frequencies and percentages. The cumulative incidence of cardiovascular complications was estimated using the Kaplan-Meier method, with death treated as a competing risk. Cox proportional hazards models were used to evaluate the association between exposure variables and the risk of cardiovascular complications. Hazard ratios (HR) and 95% CIs were calculated. To assess the impact of treatment era, analyses were stratified by decade of diagnosis (1970s, 1980s, and 1990s) and tested for trends. A risk prediction model was developed using multivariable logistic regression, internally validated using bootstrapping techniques, and its performance was assessed using the C statistic and calibration plots. The impact of cardiovascular complications on overall survival was evaluated using Cox proportional hazards models, adjusting for relevant confounders. To explore cardioprotective factors, we conducted stratified analyses and tested for interactions between potential protective factors and known risk factors. Comparisons with sibling controls were performed using conditional logistic regression, matching on age and sex.

Additional Analytical Considerations

The proportional hazards assumption for Cox regression models was tested using Schoenfeld residuals and time-dependent covariate analyses. No significant violations were identified for primary variables of interest (all $P>.10$). Missing data for covariates (primarily BMI [8.2%] and smoking status [6.5%]) were handled using multiple imputation with chained equations, generating 20 imputed datasets. Sensitivity analyses comparing complete case analyses and imputed data showed consistent results. Quality of life was assessed using the 36-Item Short Form Health Survey instrument, with particular attention to physical functioning, role limitations due to physical health, general health, and vitality domains, which showed the largest decrements among survivors with cardiovascular complications.

Sensitivity Analyses

We conducted several sensitivity analyses to assess the robustness of our findings, including multiple imputation for missing data, analyses restricted to participants with medical record-confirmed cardiovascular outcomes, and evaluation of potential selection bias due to loss to follow-up.

Ethical Considerations

This study was conducted in accordance with the Declaration of Helsinki and approved by the Ethics Committee of Healthy Steps Pediatrics (protocol code 2024-141 on October 1, 2024). This study solely used retrospective preexisting data; thus institutional review board approval was waived.

Results

Study Population

Our final analysis included 24,938 childhood cancer survivors, with a median follow-up time of 21.3 years (IQR 15.8 - 27.6). The median age at cancer diagnosis was 7.2 years (IQR 3.4 - 13.5; range 0 - 20.9), and 53.6% of the cohort was male. The most common cancer diagnoses were leukemia (34.1%), lymphoma (19.7%), and central nervous system tumors (13.2%). These statistics are presented in Table 1.

Table . Demographic and clinical characteristics of childhood cancer survivors.

Characteristic	All survivors (N=24,938)	With cardiovascular complications (n=2743)	Without cardiovascular complications (n=22,195)	*P* value
Demographic factor				
Age at diagnosis (years), median (IQR)	7.2 (3.4 - 13.5)	9.6 (5.1 - 15.2)	6.9 (3.1 - 13.1)	<.001
Sex (male), n (%)	13,367 (53.6)	1728 (63)	11,639 (52.4)	<.001
Race/ethnicity, n (%)				.08
White, non-Hispanic	20,666 (83)	2304 (84)	18,395 (82.9)	
Black, non-Hispanic	1146 (4.6)	138 (5)	1008 (4.5)	
Hispanic	1971 (7.9)	193 (7)	1778 (8)	
Other	1122 (4.5)	108 (3.9)	1014 (4.6)	
Clinical factor				
Primary diagnosis, n (%)				<.001
Leukemia	8504 (34.1)	817 (29.8)	7687 (34.6)	
Lymphoma	4913 (19.7)	662 (24.1)	4251 (19.2)	
Central nervous system tumor	3292 (13.2)	329 (12)	2963 (13.4)	
Sarcomas	3316 (13.3)	467 (17)	2849 (12.8)	
Other	4913 (19.7)	468 (17.1)	4445 (20)	
Treatment era (years), n (%)				<.001
1970 - 1979	8730 (35)	1180 (43)	7550 (34)	
1980 - 1989	9477 (38)	998 (36.4)	8479 (38.2)	
1990 - 1999	6731 (27)	565 (20.6)	6166 (27.8)	
Treatment exposure				<.001
Anthracycline exposure, n (%)	14,241 (57)	1974 (72)	12,240 (55.1)	
Chest radiation, n (%)	9726 (39)	1536 (56)	8190 (36.9)	
Current status				<.001
Age at last follow-up (years), median (IQR)	29.3 (23.5 - 37.1)	34.2 (27.8 - 41.5)	28.5 (22.9 - 36.2)	
BMI ≥30 kg/m^2, n (%)	7481 (30)	987 (36)	6494 (29.3)	
Current smoker, n (%)	3741 (15)	494 (18)	3247 (14.6)	

Incidence of Cardiovascular Complications (Objective 1)

During the follow-up period, 2743 (11%) survivors developed at least 1 cardiovascular complication. The 30-year cumulative incidence of any cardiovascular complication was 18.7% (95% CI 17.9% - 19.5%) after cancer diagnosis. Specific complication rates included cardiomyopathy 7.4% (95% CI 6.9% - 7.9%), coronary artery disease 3.8% (95% CI 3.5% - 4.1%), valvular heart disease 5.2% (95% CI 4.8% - 5.6%), and arrhythmias 6.9% (95% CI 6.4% - 7.4%) and are presented in Table 2.

Table . Summary of key findings.

Cardiovascular outcome	Cases, n (%)	Cumulative incidence at 30 years (%; 95% CI)
Any cardiovascular complication	2743 (11)	18.7 (17.9 - 19.5)
Cardiomyopathy	1845 (7.4)	7.4 (6.9 - 7.9)
Coronary artery disease	948 (3.8)	3.8 (3.5 - 4.1)
Valvular heart disease	1297 (5.2)	5.2 (4.8 - 5.6)
Arrhythmias	1721 (6.9)	6.9 (6.4 - 7.4%)

Temporal Patterns and Treatment Era Effects (Objectives 2 and 4)

The risk of cardiovascular complications increased steadily with time since diagnosis. However, we observed a significant trend of decreasing risk across treatment eras, as presented in Table 3 (*P* for trend<.001). Compared with patients treated in the 1970s, those treated in the 1990s had a 25% lower risk of developing cardiovascular complications (HR 0.75, 95% CI 0.67 - 0.84).

Table . Treatment era analysis.

Treatment era (years)	N	Events, n (%)	Cumulative incidence at 30 years (%; 95% CI)	Adjusted hazard ratio (95% CI)	*P* value
1970 - 1979	8730	1180 (13.5)	22.3 (20.9 - 23.7)	1.00 (reference)	—[a]
1980 - 1989	9477	998 (10.5)	18.1 (16.8 - 19.4)	0.83 (0.76 - 0.91)	<.001
1990 - 1999	6731	565 (8.4)	14.5 (12.7 - 16.3)	0.75 (0.67 - 0.84)	<.001
P for trend	—	—	—	—	<.001

[a]Not applicable.

Risk Factors for Cardiovascular Complications (Objective 3)

In multivariable Cox regression analyses, several factors were significantly associated with increased risk of cardiovascular complications (Table 4), such as anthracycline exposure (HR 2.31, 95% CI 2.09 - 2.55) for cumulative doses ≥250 mg/m², chest radiation (HR 1.84, 95% CI 1.66 - 2.05) for doses ≥20 Gy, age at diagnosis (per year increase; HR 1.05, 95% CI 1.03 - 1.07), male sex (HR 1.28, 95% CI 1.18 - 1.39), and BMI ≥30 kg/m² (HR 1.45, 95% CI 1.31 - 1.61).

Table . Risk factors for cardiovascular complications in childhood cancer survivors.

Risk factor	Adjusted hazard ratio (95% CI)	*P* value
Treatment factors		
Anthracycline dose		
None	1.00 (reference)	—[a]
1 - 149 mg/m^2	1.56 (1.38 - 1.76)	<.001
150 - 249 mg/m^2	1.93 (1.73 - 2.15)	<.001
≥250 mg/m^2	2.31 (2.09 - 2.55)	<.001
Chest radiation dose		
None	1.00 (reference)	—
1 - 19 Gy	1.32 (1.17 - 1.49)	<.001
≥20 Gy	1.84 (1.66 - 2.05)	<.001
Demographic factors		
Age at diagnosis (per year increase)	1.05 (1.03 - 1.07)	<.001
Sex (male)	1.28 (1.18 - 1.39)	<.001
Lifestyle/modifiable factors		
BMI		
<25 kg/m^2	1.00 (reference)	—
25 - 29.9 kg/m^2	1.21 (1.09 - 1.34)	<.001
≥30 kg/m^2	1.45 (1.31 - 1.61)	<.001
Current smoking	1.33 (1.20 - 1.47)	<.001
Physical inactivity	1.19 (1.08 - 1.31)	<.001
Medical comorbidities		
Hypertension	1.51 (1.36 - 1.67)	<.001
Diabetes	1.47 (1.29 - 1.68)	<.001
Dyslipidemia	1.32 (1.19 - 1.47)	<.001

[a]Not applicable.

Risk Prediction Model (Objective 5)

Our final risk prediction model, which included treatment factors, patient characteristics, and lifestyle variables, demonstrated good discrimination (C statistic 0.78, 95% CI 0.76 - 0.80). For internal validation, we used bootstrapping with 1000 resamples, which confirmed the model's robustness (optimism-corrected C statistic 0.76). Calibration assessment using the Hosmer-Lemeshow goodness-of-fit test showed adequate calibration (*P*=.42).

While external validation was not feasible in this study due to the lack of comparable cohorts with similar long-term follow-up, we developed a simplified risk score system based on the model coefficients to facilitate clinical application. This scoring system assigns points to key risk factors, anthracycline dose (0 - 3 points), chest radiation dose (0 - 3 points), age at diagnosis (0 - 2 points), sex (0 - 1 point), and BMI category (0 - 2 points), with a total score range of 0 - 11. Scores ≥7 identify survivors at high risk (>25% 30-year cumulative incidence) who may benefit from enhanced cardiovascular surveillance.

Impact on Survival and Quality of Life (Objective 6)

Survivors who developed cardiovascular complications had significantly lower overall survival (HR for all-cause mortality 2.3, 95% CI 2.1 - 2.5) and reported lower quality-of-life scores across multiple domains (*P*<.001 for all comparisons).

Exploration of Cardioprotective Factors (Objective 7)

We conducted comprehensive analyses to evaluate potential cardioprotective factors among childhood cancer survivors. Several significant protective associations emerged (Textbox 2).

Textbox 2. Cardioprotective factors among childhood cancer survivors.

- Physical activity: survivors who engaged in regular physical activity (defined as ≥150 min of moderate-intensity exercise/wk) had a 20% lower risk of cardiovascular complications (hazard ratio [HR] 0.80, 95% CI 0.72 - 0.89; *P*<.001). This protective effect remained significant after adjusting for treatment exposures and demographic factors.
- Cardioprotective medications: survivors who received cardioprotective medications showed a reduced risk of cardiovascular complications, as follows: (1) angiotensin-converting enzyme inhibitors: 18% risk reduction (HR 0.82, 95% CI 0.73 - 0.91), (2) β-blockers: 15% risk reduction (HR 0.85, 95% CI 0.76-0.95), and (3) statins: 12% risk reduction (HR 0.88, 95% CI 0.79-0.98).
- Dexrazoxane administration: among patients who received anthracyclines, concurrent dexrazoxane administration was associated with a 35% lower risk of cardiomyopathy (HR 0.65, 95% CI 0.54-0.78).
- Nutritional factors: adherence to a Mediterranean diet was associated with a 16% lower risk of cardiovascular complications (HR 0.84, 95% CI 0.75-0.94).

These findings suggest multiple avenues for risk reduction through lifestyle modifications, pharmacological interventions, and treatment adaptations that may be incorporated into survivorship care plans.

Comparison With Sibling Controls (Objective 8)

As presented in Table 5, compared with sibling controls, childhood cancer survivors had a significantly higher risk of cardiovascular complications (odds ratio 3.7, 95% CI 3.2 - 4.2). This excess risk was most pronounced for cardiomyopathy (odds ratio 5.2, 95% CI 4.3 - 6.3).

Table . Comparison with sibling controls.

Cardiovascular outcome	Survivors (N=24,938), n (%)	Siblings (N=5085), n (%)	Age- and sex-adjusted odds ratio (95% CI)	Fully adjusted odds ratio (95% CI)[a]
Any cardiovascular outcome	2743 (11)	157 (3.1)	3.7 (3.2 - 4.2)	3.5 (3 - 4)
Cardiomyopathy	1845 (7.4)	73 (1.4)	5.2 (4.3 - 6.3)	4.8 (4 - 5.8)
Coronary artery disease	948 (3.8)	68 (1.3)	2.8 (2.2 - 3.5)	2.6 (2 - 3.3)
Valvular heart disease	1297 (5.2)	76 (1.5)	3.4 (2.8 - 4.1)	3.1 (2.6 - 3.7)
Arrhythmias	1721 (6.9)	70 (1.4)	3.3 (2.8 - 3.9)	3.1 (2.6 - 3.7)

[a]Adjusted for age, sex, race/ethnicity, BMI, smoking status, and family history of cardiovascular disease.

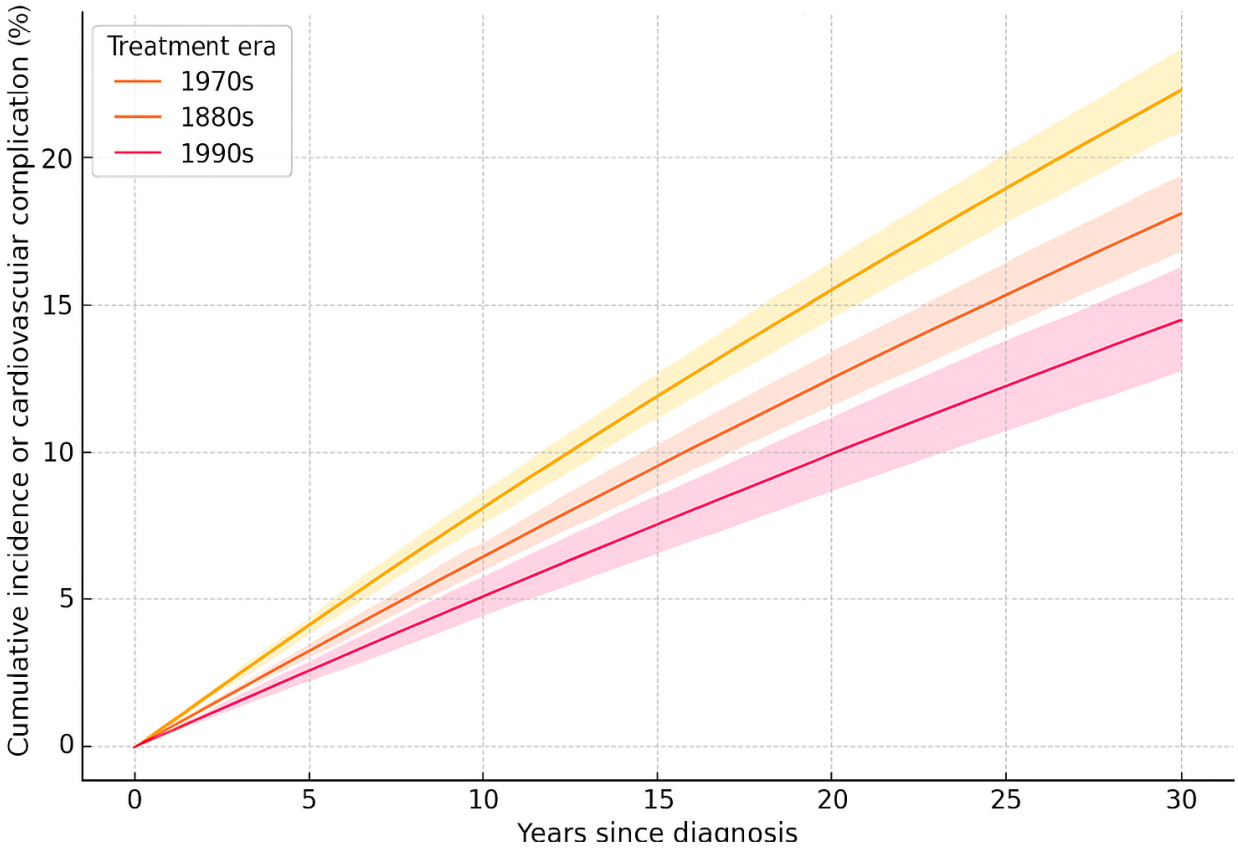
Discussion

Principal Findings

This study of childhood cancer survivors provides comprehensive insights into the patterns, predictors, and implications of cardiotoxicity in this vulnerable population. Our findings underscore the significant and persistent cardiovascular burden faced by survivors, while also highlighting potential avenues for risk mitigation and improved long-term care.

As presented in Figure 1, the cumulative incidence of cardiovascular complications in our cohort reached 18.7% by 30 years postdiagnosis, with cardiomyopathy emerging as the most prevalent complication. This incidence is substantially higher than that observed in the general population and aligns with previous studies suggesting an elevated cardiovascular risk in childhood cancer survivors [3,7]. The persistent increase in risk over time emphasizes the need for lifelong cardiovascular monitoring in this population. Kaplan-Meier–style curves in Figure 1 display the cumulative incidence (%) from diagnosis to 30 years' follow-up. Shaded bands depict the 95% CIs derived from the reported 30-year incidences (1970s: 22.3%, 1980s: 18.1%, and 1990s: 14.5%). The downward trend across eras (log-rank test, *P*<.001) illustrates the impact of evolving cardioprotective treatment protocols.

Figure 1. Cumulative incidence of any cardiovascular complication among childhood cancer survivors, stratified by treatment era (1970s vs 1980s vs 1990s).



We acknowledge that the observed trend of decreasing cardiovascular risk across treatment eras might be partially influenced by survivor selection bias. Patients with severe early toxicity resulting in mortality would be systematically excluded from later follow-up, potentially leading to an underestimation of cardiotoxicity risk. To address this concern, we conducted sensitivity analyses using inverse probability weighting to account for potentially informative censoring, which yielded similar, albeit slightly higher, risk estimates (adjusted HR for treatment in the 1990s vs 1970s: 0.79; 95% CI 0.70 - 0.89). In addition, we compared treatment protocols across eras and found that reductions in anthracycline doses and implementation of cardiac-sparing radiation techniques likely contributed to the genuine reduction in cardiovascular risk in more recent cohorts.

Our analysis confirmed several established risk factors for cardiotoxicity, including anthracycline exposure and chest radiation [4]. The dose-dependent relationship observed for both treatments reinforces the importance of treatment optimization to minimize cardiac risk without compromising oncological efficacy. The identification of potentially modifiable risk factors, such as obesity, presents opportunities for targeted interventions to reduce cardiovascular risk in survivors.

The observed higher risk of cardiovascular complications in male survivors (HR 1.28, 95% CI 1.18 - 1.39) warrants further consideration. This gender disparity may be attributed to multiple factors. First, male survivors were more likely to receive higher cumulative anthracycline doses (median 240 mg/m² vs 210 mg/m² in females; *P*<.001) and chest radiation (43% vs 35%; *P*<.001). However, the increased risk persisted after adjusting for these treatment exposures, suggesting additional mechanisms. Male survivors in our cohort also demonstrated higher rates of cardiovascular comorbidities such as dyslipidemia (26% vs 19%; *P*<.001), which may have exacerbated subclinical cardiac damage. Furthermore, biological differences in cardioprotection, particularly the role of estrogen in females, may contribute to this disparity, as observed in the general population [15].

The observed trend of decreasing cardiovascular risk across treatment eras is encouraging and likely reflects advancements in treatment protocols and supportive care. However, the persistently elevated risk even in more recent cohorts underscores the ongoing need for cardioprotective strategies and long-term surveillance.

Clinical Implications

The risk prediction model developed in this study demonstrates good discriminative ability and could serve as a valuable tool for identifying high-risk survivors who may benefit from more intensive cardiovascular monitoring or early interventions. Integrating this model into clinical practice could facilitate personalized follow-up strategies and resource allocation.

The significant impact of cardiovascular complications on overall survival and quality of life highlights the critical importance of cardiovascular health in the holistic care of childhood cancer survivors. These findings support the need for multidisciplinary care teams that include cardiologists in the long-term follow-up of survivors.

Our risk prediction model could be integrated with existing risk stratification systems, particularly those developed by the International Late Effects of Childhood Cancer Guideline Harmonization Group (IGHG). The IGHG guidelines currently classify survivors into high, moderate, and low-risk groups based primarily on anthracycline dose and chest radiation exposure [16]. Our model enhances this approach by incorporating additional patient factors (age, sex, and BMI) and quantifying their relative contributions to risk. A potential implementation strategy would involve using the IGHG framework for initial risk stratification, followed by our prediction model for refined risk assessment within each category. This 2-step approach would maintain consistency with established guidelines while providing more personalized risk estimates to guide surveillance frequency and intensity. Future validation studies in external cohorts could evaluate the combined performance of these complementary risk stratification approaches.

Strengths and Limitations

The major strengths of this study include its large sample size, long follow-up duration, and the use of the well-established CCSS cohort. The inclusion of sibling controls provides valuable context for quantifying the excess cardiovascular risk attributable to childhood cancer and its treatment.

However, several limitations should be considered. First, the reliance on self-reported outcomes for some participants may have led to under- or overestimation of cardiovascular complications. Specifically, self-reported data accounted for 73% of cardiovascular events, representing a limitation despite the high confirmation rate (93%) observed in the validated subset. To minimize potential reporting bias, we conducted sensitivity analyses restricted to medically confirmed cases, which yielded similar results. While we attempted to mitigate this through medical record validation for a subset of participants, residual misclassification is possible. Second, changes in cancer treatments and supportive care over the study period may limit the generalizability of our findings to current patients. Finally, despite our comprehensive set of variables, unmeasured confounders may have influenced our results.

Our study focused on clinically evident cardiovascular complications and did not assess subclinical cardiotoxicity, which might be detected through cardiac biomarkers (eg, troponins and N-terminal pro-B-type natriuretic peptide) or advanced imaging techniques (eg, echocardiography and cardiac magnetic resonance imaging). The prevalence of subclinical cardiac dysfunction is likely higher than the reported clinically apparent complications. Future studies incorporating these assessment modalities would enable earlier detection of cardiac damage and potentially identify opportunities for preventive interventions before clinical manifestation.

Future Directions

This study lays the groundwork for several important avenues of future research, including prospective studies incorporating advanced cardiac imaging and biomarkers to detect subclinical cardiac dysfunction in survivors, investigation of genetic factors that may modulate individual susceptibility to treatment-related cardiotoxicity, randomized controlled trials of cardioprotective interventions in high-risk survivors, long-term follow-up studies of more contemporary cohorts to assess the impact of modern treatment protocols on cardiovascular outcomes, and implementation studies to evaluate the clinical utility and cost-effectiveness of risk-based screening strategies.

Conclusion

Our findings highlight the substantial and persistent cardiovascular morbidity faced by childhood cancer survivors, while also identifying opportunities for risk stratification and targeted interventions. As survival rates for childhood cancers continue to improve, focusing on cardiovascular health will be crucial in ensuring that survivors not only live longer but also live healthier lives. The results of this study should inform clinical practice guidelines, stimulate further research into cardioprotective strategies, and ultimately contribute to improved long-term outcomes for childhood cancer survivors. The risk prediction model and identified protective factors provide valuable tools for refined risk stratification and targeted interventions.

Authors' Contributions

MM led the conceptualization and project administration, with AI contributing equally to both. MM was responsible for data curation, investigation, methodology, resources, supervision, and validation. Formal analysis and visualization were led by MM with supporting contributions from AI. MM prepared the original draft with support from AI. Both authors contributed to the review and editing of the manuscript.

Conflicts of Interest

None declared.

References

1. Siegel RL, Miller KD, Wagle NS, Jemal A. Cancer statistics, 2023. CA Cancer J Clin 2023 Jan;73(1):17-48. [doi: 10.3322/caac.21763] [Medline: 36633525]
2. Armstrong GT, Kawashima T, Leisenring W, et al. Aging and risk of severe, disabling, life-threatening, and fatal events in the childhood cancer survivor study. J Clin Oncol 2014 Apr 20;32(12):1218-1227. [doi: 10.1200/JCO.2013.51.1055] [Medline: 24638000]

3. Lipshultz SE, Adams MJ, Colan SD, et al. Long-term cardiovascular toxicity in children, adolescents, and young adults who receive cancer therapy: pathophysiology, course, monitoring, management, prevention, and research directions: a scientific statement from the American Heart Association. Circulation 2013 Oct 22;128(17):1927-1995. [doi: 10.1161/CIR.0b013e3182a88099] [Medline: 24081971]
4. Armenian SH, Hudson MM, Mulder RL, et al. Recommendations for cardiomyopathy surveillance for survivors of childhood cancer: a report from the International Late Effects of Childhood Cancer Guideline Harmonization Group. Lancet Oncol 2015 Mar;16(3):e123-e136. [doi: 10.1016/S1470-2045(14)70409-7] [Medline: 25752563]
5. Volkova M, Russell R. Anthracycline cardiotoxicity: prevalence, pathogenesis and treatment. Curr Cardiol Rev 2011 Nov;7(4):214-220. [doi: 10.2174/157340311799960645] [Medline: 22758622]
6. Trachtenberg BH, Landy DC, Franco VI, et al. Anthracycline-associated cardiotoxicity in survivors of childhood cancer. Pediatr Cardiol 2011 Mar;32(3):342-353. [doi: 10.1007/s00246-010-9878-3] [Medline: 21221562]
7. Mulrooney DA, Yeazel MW, Kawashima T, et al. Cardiac outcomes in a cohort of adult survivors of childhood and adolescent cancer: retrospective analysis of the Childhood Cancer Survivor Study cohort. BMJ 2009 Dec 8;339:b4606. [doi: 10.1136/bmj.b4606] [Medline: 19996459]
8. Lipshultz SE, Landy DC, Lopez-Mitnik G, et al. Cardiovascular status of childhood cancer survivors exposed and unexposed to cardiotoxic therapy. J Clin Oncol 2012 Apr 1;30(10):1050-1057. [doi: 10.1200/JCO.2010.33.7907] [Medline: 22393080]
9. Armenian SH, Sun CL, Vase T, et al. Cardiovascular risk factors in hematopoietic cell transplantation survivors: role in development of subsequent cardiovascular disease. Blood 2012 Nov 29;120(23):4505-4512. [doi: 10.1182/blood-2012-06-437178] [Medline: 23034279]
10. van Dalen EC, Caron HN, Dickinson HO, Kremer LC. Cardioprotective interventions for cancer patients receiving anthracyclines. Cochrane Database Syst Rev 2011 Jun 15;2011(6):CD003917. [doi: 10.1002/14651858.CD003917.pub4] [Medline: 21678342]
11. Oeffinger KC, Mertens AC, Sklar CA, et al. Chronic health conditions in adult survivors of childhood cancer. N Engl J Med 2006 Oct 12;355(15):1572-1582. [doi: 10.1056/NEJMsa060185] [Medline: 17035650]
12. Robison LL, Armstrong GT, Boice JD, et al. The Childhood Cancer Survivor Study: a National Cancer Institute-supported resource for outcome and intervention research. J Clin Oncol 2009 May 10;27(14):2308-2318. [doi: 10.1200/JCO.2009.22.3339] [Medline: 19364948]
13. Hudson MM, Ness KK, Gurney JG, et al. Clinical ascertainment of health outcomes among adults treated for childhood cancer. JAMA 2013 Jun 12;309(22):2371-2381. [doi: 10.1001/jama.2013.6296] [Medline: 23757085]
14. Armstrong GT, Oeffinger KC, Chen Y, et al. Modifiable risk factors and major cardiac events among adult survivors of childhood cancer. J Clin Oncol 2013 Oct 10;31(29):3673-3680. [doi: 10.1200/JCO.2013.49.3205] [Medline: 24002505]
15. Qian C, Liu J, Liu H. Targeting estrogen receptor signaling for treating heart failure. Heart Fail Rev 2024 Jan;29(1):125-131. [doi: 10.1007/s10741-023-10356-9] [Medline: 37783987]
16. de Baat EC, van Dalen EC, Mulder RL, et al. Primary cardioprotection with dexrazoxane in patients with childhood cancer who are expected to receive anthracyclines: recommendations from the International Late Effects of Childhood Cancer Guideline Harmonization Group. Lancet Child Adolesc Health 2022 Dec;6(12):885-894. [doi: 10.1016/S2352-4642(22)00239-5]

Abbreviations

CCSS: Childhood Cancer Survivor Study
HR: hazard ratio
IGHC: International Late Effects of Childhood Cancer Guideline Harmonization Group

Edited by F Wu; submitted 12.08.24; peer-reviewed by A Adhikari, J Lucas Jr; revised version received 23.04.25; accepted 23.05.25; published 31.07.25.

Please cite as:
Mansoor M, Ibrahim A
Cardiotoxicity in Pediatric Cancer Survivorship: Retrospective Cohort Study
JMIRx Med 2025;6:e65299
URL: https://xmed.jmir.org/2025/1/e65299
doi:10.2196/65299

Predicting Escalation of Care for Childhood Pneumonia Using Machine Learning: Retrospective Analysis and Model Development

Oguzhan Serin[1], MD; Izzet Turkalp Akbasli[1], MD; Sena Bocutcu Cetin[1], MD; Busra Koseoglu[1], MD; Ahmet Fatih Deveci[2], MSc; Muhsin Zahid Ugur[2], PhD; Yasemin Ozsurekci[3], MD

[1]Department of Pediatrics, Hacettepe University Medical School, Gevher Nesibe Avenue, Altindag, Ankara, Turkey

[2]Department of Health Information Systems, University of Health Sciences, Istanbul, Turkey

[3]Department of Pediatric Infectious Diseases, Hacettepe University Medical School, Ankara, Turkey

Corresponding Author:
Izzet Turkalp Akbasli, MD
Department of Pediatrics, Hacettepe University Medical School, Gevher Nesibe Avenue, Altindag, Ankara, Turkey

Related Articles:

Companion article: https://www.medrxiv.org/content/10.1101/2024.02.22.24303209v1

Companion article: https://med.jmirx.org/2025/1/e71100

Companion article: https://med.jmirx.org/2025/1/e71369

Companion article: https://med.jmirx.org/2025/1/e71098

Abstract

Background: Pneumonia is a leading cause of mortality in children aged <5 years. While machine learning (ML) has been applied to pneumonia diagnostics, few studies have focused on predicting the need for escalation of care in pediatric cases. This study aims to develop an ML-based clinical decision support tool for predicting the need for escalation of care in community-acquired pneumonia cases.

Objective: The primary objective was to develop a robust predictive tool to help primary care physicians determine where and how a case should be managed.

Methods: Data from 437 children with community-acquired pneumonia, collected before the COVID-19 pandemic, were retrospectively analyzed. Pediatricians encoded key clinical features from unstructured medical records based on Integrated Management of Childhood Illness guidelines. After preprocessing with Synthetic Minority Oversampling Technique–Tomek to handle imbalanced data, feature selection was performed using Shapley additive explanations values. The model was optimized through hyperparameter tuning and ensembling. The primary outcome was the level of care severity, defined as the need for referral to a tertiary care unit for intensive care or respiratory support.

Results: A total of 437 cases were analyzed, and the optimized models predicted the need for transfer to a higher level of care with an accuracy of 77% to 88%, achieving an area under the receiver operator characteristic curve of 0.88 and an area under the precision-recall curve of 0.96. Shapley additive explanations value analysis identified hypoxia, respiratory distress, age, weight-for-age *z* score, and complaint duration as the most important clinical predictors independent of laboratory diagnostics.

Conclusions: This study demonstrates the feasibility of applying ML techniques to create a prognostic care decision tool for childhood pneumonia. It provides early identification of cases requiring escalation of care by combining foundational clinical skills with data science methods.

(JMIRx Med 2025;6:e57719) doi:10.2196/57719

KEYWORDS

childhood pneumonia; community-acquired pneumonia; machine learning; clinical decision support system; prognostic care decision

Introduction

Pneumonia is responsible for 14% of all mortality in children aged <5 years and is included in World Health Organization (WHO) reports as the cause of death in 740,180 children in 2019 alone [1,2]. The Global Action Plan for the Prevention and Control of Pneumonia and Diarrhea, which was released by the WHO and UNICEF, aimed to reduce the mortality rate from pneumonia and diarrhea in children aged <5 years [2,3]. They have set targets that include vaccination, water and air sanitation, exclusively breastfeeding in the first 6 months, and eliminating pediatric HIV cases, along with appropriate pneumonia and diarrhea care.

It has been demonstrated that timely and accurate diagnosis of pneumonia and appropriately initiated treatment reduce mortality by up to 28% [4]. Diagnosis can often be difficult, since the clinical presentation of pneumonia in children is variable [5]. For this reason, the WHO has published the Integrated Management of Childhood Illness (IMCI) guidelines, which guide physicians in diagnosing, treating, and identifying danger signs of pneumonia [6]. While some cases of pneumonia are treatable with appropriate interventions, even low-cost or low-tech options [1], pneumonia remains a leading cause of morbidity and mortality, particularly in resource-limited countries and regions [2]. Managing high-risk populations continues to present significant challenges, especially in intensive care settings where patients often require advanced respiratory support. In addition, it has been shown that families seeking health services in resource-limited settings causes delays in providing appropriate treatment, leading to disease progression [7]. These highlight the need to improve medical care decisions, particularly in regions with limited resources, to reduce pneumonia-related morbidity and mortality.

Early and accurate recognition of patients who may require escalation of care to tertiary facilities is essential, particularly for those who will require mechanical ventilation or advanced respiratory support [8]. Predicting which patients will deteriorate is challenging due to the heterogeneous presentation of pneumonia, and clinical features such as hypoxia, respiratory distress, nutritional status, and comorbidities are critical markers that necessitate closer monitoring or transfer [9,10]. Prolonged duration of illness and failure to respond to initial treatments are also important as they may indicate inadequate treatment, misdiagnosis, or incorrect identification of potential pathogens, which can lead to the escalation of care [7,11].

Data science can provide actionable evidence for effective clinical intervention in pediatric diseases in the future [12] and can reduce inequality in health care [13]. Also, using big data and machine learning (ML) technologies is promising for childhood pneumonia in low- and middle-income countries (LMICs), especially patient-risk stratification for developing severe disease and mortality [14]. Because of their flexibility and high accuracy, ML models are used in medicine in the fields of prediction (prognostics) and classification (diagnostics) [12]. Additionally, the use of ML offers great promise for decision support in managing community-acquired pneumonia (CAP) in children, as demonstrated in recent studies. These include predicting intensive care unit needs [15], low-cost and noninvasive diagnostics for childhood pneumonia in resource-limited settings [16], supporting pathogen identification at admission only using basic clinical and laboratory features [11], and using natural language processing with ML for supporting clinical decisions on radiology reports [17].

It has been seen that the vast majority of data science studies on pneumonia aims to provide diagnostic support to the physician by processing radiological images [18]. However, diagnostic utilities are mostly unavailable in LMICs and primary care units. Therefore, physicians need prognostic support algorithms that distinguish between serious and nonserious cases without using advanced diagnostic equipment.

We aimed to develop an ML-based clinical decision support tool for childhood pneumonia that can be used by non–intensive care physicians, particularly those working in LMICs, in predicting the escalation of care and thereby ensuring the effective diagnosis and treatment of pneumonia, which is one of the 2025 goals of the WHO [1,3].

Methods

Case Definition and Patient Selection

Our study included pediatric patients who received inpatient treatment at Hacettepe University Medical School, a large, urban, tertiary, academic medical center in Ankara, Türkiye, between January 2014 and April 2020. The center serves a diverse range of pediatric patients from both urban and rural areas across the country, including those requiring advanced multidisciplinary care as well as those with less severe conditions. All patients were diagnosed with CAP based on the most recent IMCI guidelines, which provide a structured clinical framework focused on clinical features rather than advanced imaging or laboratory results [6,19]. Patients younger than 28 days of age (neonatal age), those older than 18 years, and those who had been hospitalized within the last 14 days were excluded.

The medical records of 437 patients were retrospectively examined by pediatricians, who encoded the candidate features from unstructured admission notes based on the IMCI guidelines (Tables 1 and 2). These variables were chosen based on their clinical value in clinical decision-making and their availability in primary care.

Table . Candidate features: clinical variables.

Clinical variables	Description
Age	Age in months at the time of admission
Weight (*z* score)	Standardized score based on Turkish children reference values [20], indirectly reflecting nutritional status
Gender	Biological sex (male or female)
Complaint period	Duration (days) from symptom onset to admission
Comorbidity	Presence of any significant underlying medical conditions, including congenital disorders, genetic syndromes, neuromuscular diseases, and chronic respiratory or cardiac issues
Recent antibiotics usage	Prescribed oral antibiotic use within the 14 days before admission, suggesting an inadequately treated infection or failure to respond initial care
Fever	Presence of elevated body temperature at admission
Cough	A key respiratory symptom at admission
Loss of appetite	Sign of systemic illness, reflecting impact on the patient's well-being
Respiratory distress	Presence of shortness of breath, rapid breathing (tachypnea), nasal flaring, or chest wall retractions at initial examination
Abnormal lung sounds	Auscultatory findings (eg, crackles or wheezing), indicative of pulmonary pathology at initial examination
Hypoxia	SaO_2[a] measured by pulse oximetry; hypoxia is defined as SaO_2 below 92% at initial examination
Level of care severity	Primary outcome; whether the patient requires pneumonia care at a tertiary care unit, including PICU[b] admission or respiratory support (oxygenation or ventilation), at any point during the hospital stay

[a]SaO_2: peripheral blood oxygen saturation.

[b]PICU: pediatric intensive care unit.

Table . Candidate features: laboratory variables.

Laboratory variables	Unit
Hemoglobin	Grams per deciliter (g/dL)
Leukocytes	Cells per liter ($\times 10^6$/L)
Lymphocytes	Cells per liter ($\times 10^6$/L)
Neutrophils	Cells per liter ($\times 10^6$/L)
Platelets	Cells per liter ($\times 10^9$/L)
C-reactive protein	Milligrams per liter (mg/L)
Albumin	Grams per deciliter (g/dL)
Sodium	Milliequivalents per liter (mEq/L)
Aspartate aminotransferase	Units per liter (U/L)
Alanine aminotransferase	Units per liter (U/L)

The primary outcome was the "level of care severity," scaled as severe or nonsevere. This categorization was made by physician-encoders based on whether the patient required referral to a tertiary care unit, using medical notes during the hospital stay. Children classified as severe included those admitted to the pediatric intensive care unit or those who required oxygenation or ventilation support at any time during the hospital stay.

Ethical Considerations

This study's design and procedures were approved by the Hacettepe University Clinical Research Ethics Committee with protocol GO-20/1182. Since this study is a retrospective analysis using previously collected data, informed consent was not required as per the ethics committee's approval. All data used in this study were deidentified before analysis to ensure participant privacy and confidentiality. No compensation was

provided to participants, as this study did not involve direct human participant recruitment.

Study Population

This study included 437 hospitalized patients with CAP, categorized into nonsevere (n=133, 30.4%) and severe cases (n=304, 69.6%). Demographic and clinical candidate variables, along with laboratory indices, were collected. Group comparisons were made using the Mann-Whitney *U* test for continuous variables and the χ^2 test for categorical variables, with significance set at $P<.05$. A summary of these characteristics and statistical comparisons are provided in Table 3.

Table . Characteristics of the study population by level of care severity (N=437).

Candidate variables	Nonsevere (n=133, 30.4%)	Severe (n=304, 69.6%)	Test statistic (*df*)	*P* value
Age (months), median (IQR)	44 (13 to 98)	23 (7 to 64.5)	16,602[a]	.003
Weight (*z* scores), median (IQR)	−0.57 (−1.4 to 0.45)	−0.7 (−2.5 to 0.4)	17,784[a]	.045
Complaint period (days), median (IQR)	4 (2 to 7)	4 (2 to 7)	19,274[a]	.44
Gender, n (%)			0.05[a]	.83
Male	68 (30.9)	152 (69.1)		
Female	65 (30)	152 (70)		
Comorbidity, n (%)	85 (28.7)	211 (71.3)	1.28[b] (1)	.26
Recent antibiotic usage, n (%)	40 (26.3)	112 (73.7)	1.87[b] (1)	.17
Fever, n (%)	100 (32.3)	210 (67.7)	1.68[b] (1)	.20
Cough, n (%)	115 (31.3)	253 (68.8)	0.50[b] (1)	.48
Loss of appetite, n (%)	37 (32)	80 (68)	0.11[b] (1)	.74
Respiratory distress, n (%)	43 (17.1)	208 (82.9)	49.30[b] (1)	<.001
Abnormal lung sounds, n (%)	102 (26.9)	277 (73.1)	16.70[b] (1)	<.001
Hypoxia, n (%)	20 (7.7)	240 (92.3)	156.82[b] (1)	<.001
Hemoglobin (g/dL), median (IQR)	11.6 (10.4 to 12.9)	11.6 (10.6 to 12.6)	20,022[a]	.87
Leukocytes ($\times 10^6$/L), median (IQR)	9900 (6800 to 14,600)	10,950 (8050 to 15,850)	17,837[a]	.05
Lymphocytes ($\times 10^6$/L), median (IQR)	2300 (1400 to 3700)	2800 (1900 to 4400)	17,039[a]	.01
Neutrophils ($\times 10^6$/L), median (IQR)	5285 (2700 to 9200)	6500 (3650 to 10,900)	17,645[a]	.045
Platelets ($\times 10^9$/L), median (IQR)	310 (225 to 386)	317.5 (230.5 to 425)	19,399[a]	.50
C-reactive protein (mg/L), median (IQR)	2.06 (0.79 to 7.67)	2.06 (0.83 to 7.35)	19,842[a]	.76
Albumin (g/dL), median (IQR)	3.9 (3.73 to 4.2)	3.9 (3.4 to 4.2)	17,121[a]	.01
Sodium (mEq/L), median (IQR)	136 (135 to 138)	136 (134 to 138)	19,657[a]	.64
Aspartate aminotransferase (U/L), median (IQR)	35 (26 to 42)	35 (28 to 50)	18,382[a]	.13
Alanine aminotransferase (U/L), median (IQR)	17 (12 to 26)	18 (13 to 29)	18,457[a]	.15

[a]Mann-Whitney *U* test.

[b]Chi-square test.

Data Preprocessing

Data preprocessing, analysis, visualization, and model setup were conducted using Python (version 3.12; Python Software Foundation). We used Python libraries such as *Pandas*, *NumPy*, *Matplotlib*, *Seaborn*, and *Plotly* for exploratory data analysis. For model development, the *PyCaret* library was used, which includes an unsupervised anomaly detection module to identify and handle anomalous data points. *PyCaret* also offers various preprocessing modules to iteratively handle missing data using the light gradient boosting machine (LightGBM) algorithm. In this method, missing values were treated as dependent variables

and predicted based on other available features, minimizing bias. Individual feature weights were applied during this process. Specifically, of the 415 cases, the following features had missing values: C-reactive protein (n=34, 8.2%), albumin (n=10, 2.4%), sodium (n=8, 1.9%), aspartate aminotransferase (n=16, 3.9%), and alanine aminotransferase (n=16, 3.9%). For numerical data, min-max scaling was applied, while categorical data were processed using one-hot encoding. These preprocessing steps ensured the dataset was well prepared for model training and validation.

Handling the Imbalanced Dataset

The balance of the dataset was assessed using Shannon entropy, yielding a value of 0.7, which indicates an imbalanced dataset. To address this, we applied Synthetic Minority Oversampling Technique (SMOTE)–Tomek, a refined variation of the widely recognized SMOTE. This approach combines oversampling of the minority class with the removal of overlapping samples from the majority class through Tomek links. So, the ratio of samples becomes 1:1. The *Imblearn* library was used for implementing data oversampling.

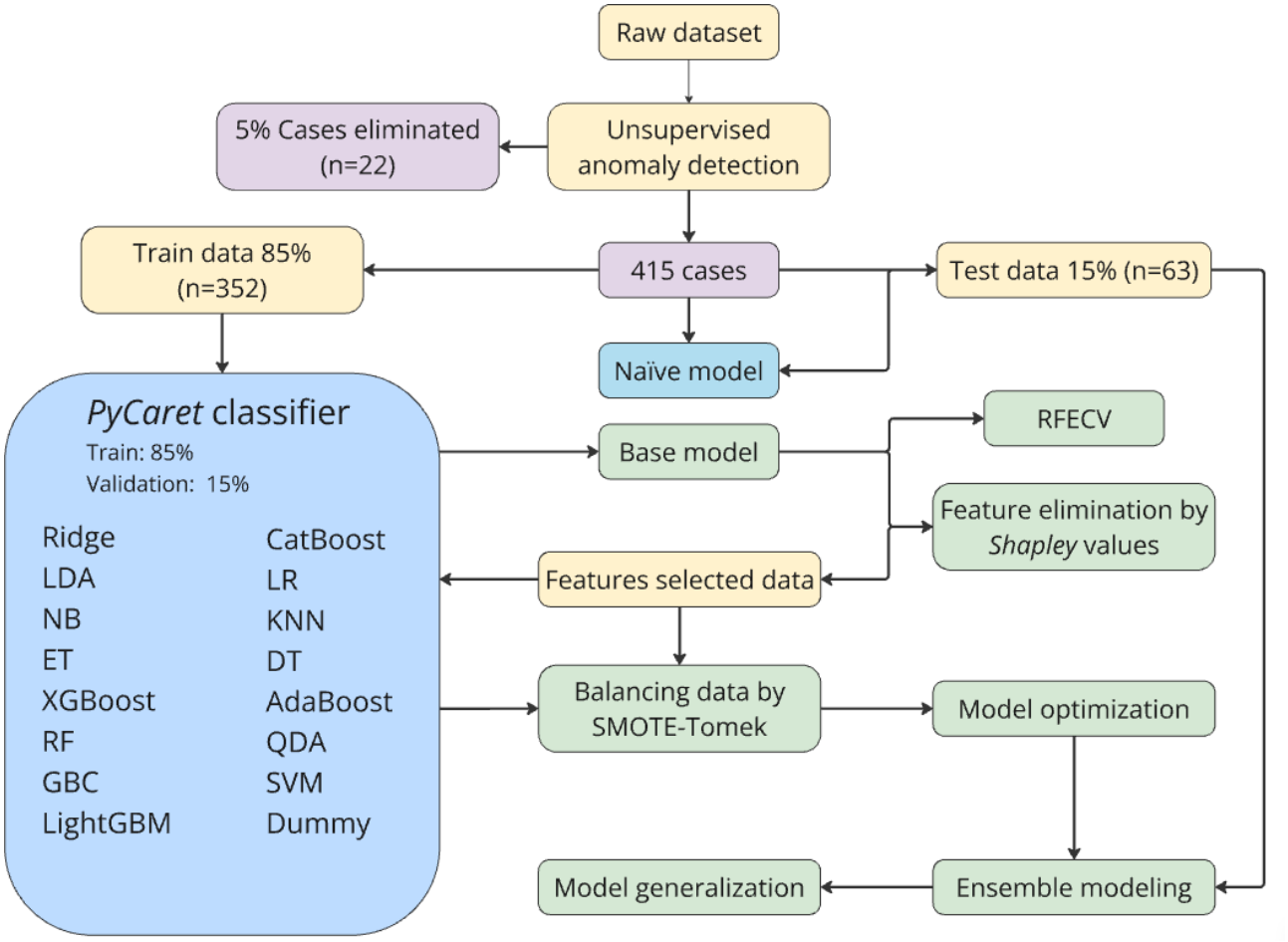
The dataset was split into two sets using the *train_test_split* method of the *SciKit-Learn* library. In the beginning, we allocated 5% of the general dataset as test data in order to prevent data leakage. The remaining 95% was split into training (352/415, 85%) and validation (63/415, 15%) sets.

Algorithms

PyCaret provides efficient implementations of state-of-the-art algorithms and is reusable among scientific disciplines. We used the *PyCaret* classifier module for classification, which includes the following models: ridge classifier, linear discriminant analysis, naïve Bayes, extra tree classifier, extreme gradient boosting (XGBoost), random forest, gradient boosting classifier, LightGBM, CatBoost classifier, logistic regression, k-neighbors classifier, decision tree, AdaBoost classifier, quadratic discriminant analysis, support vector machine with linear kernel, and dummy classifier.

In our work, we considered 10-fold cross-validation. While developing our model with *PyCaret* tools, we implemented the tuning function using the *Tune-Sklearn* library and the *hyper-band* optimization algorithm to obtain a set of best-performing parameters. For ensembling, we also used *PyCaret* classifier ensemble, stack, and blender methods. Ensembling methods have strong evidence that they can significantly enhance the accuracy of classifications [21]. After the optimization of parameters, in the last phase, we used the most common ensemble methods provided by the *PyCaret* library to further improve our model's performance (Figure 1).

Figure 1. The experimental setup: in this figure, we illustrate the experimental process of our models. Initially, we cleaned the data by identifying 5% of cases as abnormal data using unsupervised learning. We then split the data into a train set (85%) and a validation set (15%) using the *PyCaret* classifier model. The base model with the highest AUC-ROC value was the RF algorithm. Subsequently, we determined the optimal number of features as 18 using RFECV and selected the top 18 features based on Shapley values. We then balanced the dataset using the SMOTE-Tomek method and developed high-performing models. After optimizing the hyperparameters, we selected the best-performing model and created new models by using ensemble methods. In parallel, we developed a new model using only clinical findings for clinical prediction. AdaBoost: AdaBoost classifier; AUC-ROC: area under the receiver operator characteristic curve; CatBoost: CatBoost classifier; DT: decision tree; Dummy: dummy classifier; ET: extra tree classifier; GBC: gradient boosting classifier; KNN: k-neighbors classifier; LDA: linear discriminant analysis; LightGBM: light gradient boosting machine; LR: logistic regression; NB: naïve Bayes; QDA: quadratic discriminant analysis; RF: random forest; RFECV: recursive feature elimination with cross-validation; Ridge: ridge classifier; SMOTE: Synthetic Minority Oversampling Technique; SVM: support vector machine linear kernel classifier; XGBoost: extreme gradient boosting.



Feature Selection and Data-Reducing Methods

Feature selection is a process of one-by-one evaluation to determine which features are effective on the results within the dataset. Irrelevant or partially relevant features can negatively impact ML model performance and make the ML model learn based on irrelevant features. These methods are aimed at eliminating irrelevant features and keeping the strong features to reduce the dimension of the dataset. Recursive feature elimination is a feature selection method that fits a model and removes the irrelevant features until the specified number of features is reached. Recursive feature elimination with cross-validation (RFECV) aims to select the optimal number of features using permutation importance and recursive feature elimination. In this study, we used the *RFECV* module from *yellowbrick* library for selecting the optimum feature number. The Shapley additive explanations (SHAP) method is an innovative tool for explaining ML decision-making processes for datasets. The goal of the SHAP method is to present and explain the prediction with respect to the contribution of each feature to the predicted value. In RFECV, the features are ranked by a permutation importance measure. The SHAP algorithm was used for feature selection (Figure 2), as it provides more consistent and accurate importance values compared to the permutation approach. Ultimately, RFECV algorithms showed that 18 parameters are sufficient to explain nearly 90% of variances. Overall, 13 clinical and 5 laboratory variables were selected according to their SHAP values (Figure 2).

Figure 2. Feature selection: SHAP values are presented for the random forest classifier model with the highest AUC-ROC score in the dataset before feature selection, using the *SHAP* library's *plot_summary* module. The y-axis shows the importance of each feature, with the most important feature at the top and the least important at the bottom. The colors represent the contribution of each feature to the model's prediction. For example, features that have a large positive contribution to the prediction are shown in a warm color (eg, red), while features that have a large negative contribution are shown in a cool color (eg, blue). In this example, hypoxia is the most important attribute in the plot. The presence of hypoxia (hypoxia=1) causes the model to move closer to the target class, while its absence causes the model to move away from the target class. This predicts that hypoxia is an aggravating factor, while high levels of albumin have a protective effect for the target class. In summary, hypoxia is an adverse factor, and high albumin levels are protective. ALT: alanine aminotransferase; AST: aspartate aminotransferase; AUC-ROC: area under the receiver operator characteristic curve; CRP: C-reactive protein; SHAP: Shapley additive explanations.

Results

Study Population Characteristics

A comparison of the demographic and clinical characteristics between the nonsevere and severe groups is presented in Table 3. Of the 437 patients, 304 (69.6%) met the primary outcome, requiring the escalation of care. Patients in the severe care group were significantly younger, with a median age of 23 months compared to 44 months in the nonsevere level of care group (P=.003). Additionally, the severe group had lower weight z scores (P=.045).

Key clinical differences included higher rates of respiratory distress (208/304, 82.9% vs 43/133, 17.1%; P<.001), abnormal lung sounds (277/304, 73.1% vs 102/133, 26.9%; P<.001), and hypoxia (240/304, 92.3% vs 20/133, 7.7%; P<.001) in the severe group. In terms of laboratory findings, the severe group had higher leukocyte counts (P=.005), neutrophil counts (P=.045), and lymphocyte counts (P=.001). Albumin levels were slightly lower in the severe group (P=.01). No significant differences were observed between the groups in gender distribution (P=.83), comorbidities (P=.26), recent antibiotic use (P=.17), or C-reactive protein levels (P=.76).

Model Performances

In this section, we present a comparison of the performance of 16 different algorithms for raw and preprocessed datasets. We used various evaluation metrics such as accuracy, area under the receiver operator characteristic curve (AUC-ROC), recall, precision, F_1-score, Cohen κ, and Matthews correlation coefficient to assess model performance. To analyze model performance, all prediction experiments were conducted using 10-fold cross-validation. Subsequently, the models were optimized, and their performances were evaluated on a balanced dataset using SMOTE-Tomek and feature selection. The performances of the three models with the highest performance (CatBoost, XGBoost, and LightGBM) were evaluated by applying hyperparameter optimization and ensemble methods. Table 4 compares the results obtained with CatBoost, XGBoost, and LightGBM among the optimized and nonoptimized results, as well as the results of the combinations with the highest performance from the basic ensembling methods (ensembling, blending, and stacking methods). The highest AUC-ROC value was achieved by using optimized LightGBM as the meta-model in the stacking method.

Table . Comparative performance of machine learning models for the escalation of care prediction. Italicized values represent the highest scores for each column.

Model	Accuracy	AUC-ROC[a]	AUC-PRC[b]	Recall	Precision	F_1-score	Cohen κ	MCC[c]
CatBoost[d]	0.77	0.85	0.94	0.75	0.91	0.82	0.52	0.54
LightGBM[e,f]	0.80	0.87	0.96	0.79	0.92	0.85	0.58	0.59
XGBoost[f,g]	0.77	0.83	0.96	0.72	*0.94*	0.82	0.54	0.57
Ensembling[h]	0.77	0.86	0.95	0.72	*0.94*	0.82	0.54	0.57
Stacking[i]	0.80	*0.88*	0.96	0.79	0.92	0.85	0.58	0.59
Blending-1[j]	0.77	0.86	0.96	0.75	0.91	0.82	0.52	0.57
Blending-2[k]	*0.85*	0.84	0.96	*0.95*	0.85	*0.90*	*0.63*	*0.64*

[a]AUC-ROC: area under the receiver operating characteristic curve.

[b]AUC-PRC: area under the precision-recall curve.

[c]MCC: Matthews correlation coefficient.

[d]The performance of unoptimized CatBoost.

[e]LightGBM: light gradient boosting machine.

[f]The performance values obtained after optimization of XGBoost and LightGBM.

[g]XGBoost: extreme gradient boosting.

[h]The performance of the optimized LightGBM ensembling method, which achieved the highest results among CatBoost, XGBoost, and LightGBM algorithms.

[i]The performance of the model with optimized LightGBM as a meta-model in the stacking method, as it showed the highest performance.

[j]The combination of optimized LightGBM and XGBoost with higher performance in the blending method.

[k]Using the top-5, highest-ranked clinical features, the peak performance was realized by using a method that incorporated the optimized CatBoost, LightGBM, and XGBoost models.

In addition to the metrics reported in Table 4, we evaluated the performance of the *Blending-2* model using the precision-recall curve metric, which is particularly useful for imbalanced datasets. The precision-recall curve plot for this model, using the top-5 ranked clinical features, is provided in Multimedia Appendix 1. The model achieved a strong average precision-recall score of 0.96, further highlighting its robustness in handling imbalanced data.

Feature Importance

The optimized LightGBM in the model, developed with balanced and feature-selected data, was responsible for the

attainment of the highest performance. Upon evaluation of clinical features according to SHAP values, a ranking was established based on their feature importance scores, with the highest score being garnered by the top-5 clinical features (hypoxia, respiratory distress, age, *z* score of weight for age, and antibiotic usage before admission; Multimedia Appendix 2). The application of a workflow using these 5 features, as done previously, resulted in the highest accuracy performance (84%), which was achieved through the use of the ensemble method, incorporating the blending method of the optimized CatBoost, LightGBM, and XGBoost models.

Discussion

Pneumonia, the leading cause of childhood mortality, is also one of the most common causes of hospitalization [3,22]. It remains a significant global health burden, particularly in children aged <5 years, where timely and accurate clinical management is crucial for reducing mortality [8]. While prevention strategies are well documented, the clinical challenge lies in efficiently identifying patients who require escalated care. In this study, we present a contemporary approach to building an ML-based, prognostic care referral decision support tool that assists primary care physicians in determining where the case should be managed with an accuracy of more than 80%.

Today, there is widespread knowledge of the prevention, diagnosis, treatment, and management of complications in CAP, but due to resource limitations, it is not possible for all physicians and patients to benefit from this [14]. Recent advancements in medical informatics have the potential to reduce health care disparities and empower physicians in resource-limited settings [11-15], offering new hope for identifying high-risk populations and preventing mortality where current methods fall short.

The recent COVID-19 pandemic has impacted several medical fields, including the disruption of research practices by shifting researchers' focus and patient recruitment [23,24] and significantly reducing the incidence of non–COVID-19 pneumonia by preventing transmission [25-27]. In the current postpandemic state, non–COVID-19 childhood pneumonia remains a global health concern, especially in resource-limited settings according to the most recent reports [2], with respiratory infections likely to rise again as pandemic measures have already been eased [28]. Now, focusing back to reducing the mortality of CAP is critical to ensure pediatric pneumonia care benefits from recent advancements that COVID-19 provided [29,30]. This study, built primarily on prepandemic cases, provides a foundational context for future studies on CAP using ML in the postpandemic era.

Since March 2020, a substantial amount of data about COVID-19 have been published, including COVID-19–related artificial intelligence studies focused on pneumonia diagnosis by radiological findings [31]. However, pneumonia diagnosis is clinical, and routine chest radiographs are not necessary for the confirmation diagnosis [32] and do not improve outcomes [33]. In addition, chest radiography can be used only in inpatient settings to identify complications or evaluate response to treatment.

Although strong diagnostic support algorithms have been published in pneumonia-related studies in recent years, there is still a need for prognostic studies for pneumonia management [31]. Determining the severity of a disease or predicting its prognosis answers essential questions of physicians in medical decision-making, such as "Where should it be treated? Outpatient? ICU?" "Which therapy should I start? How long should I give it?" and "When should I discharge the patient? When should I call for control?" There are several studies and guidelines in the literature for severity assessment and prognosis prediction of pneumonia [9,10,34]. For the majority, mortality and the development of complications were the primary outcomes, and clinical, radiological, and laboratory variables are the key predictors. Yet, there is a limited number of studies predicting required referral to tertiary care based on basic clinical and laboratory features available in primary care settings [15].

This study reviewed important pneumonia prognostic predictors of children hospitalized in a major academic medical center. The primary outcome of interest was the level of care severity, classified as severe or nonsevere based on the need for pediatric intensive care unit admission or oxygen/ventilation support. The main objective of this study was not only to build the best model but also to answer the primary care physician's question: "Where should the case be managed?" Our model demonstrated promising predictive accuracy, with an AUC-ROC exceeding 0.85 and an accuracy of 77% to 88% (Table 4). The key clinical features identified—hypoxia, respiratory distress, age, *z* score of weight for age, and complaint period (Multimedia Appendix 2)—align with existing clinical guidelines, which emphasize the importance of respiratory and nutritional status in predicting disease severity [33-36].

In this study, we used SMOTE-Tomek, a method proven effective in medical tasks, to address class imbalance without losing valuable clinical information [37,38], which was essential given the significantly imbalanced and small sample–sized dataset. Additionally, we used RFECV and SHAP, both of which have been established as robust methods in previous studies [11,39,40], for feature selection. These techniques not only improved our model's performance but also allowed us to isolate the most clinically significant features (Figure 2, also see Multimedia Appendix 2), enabling clinicians to decide using their own skills without involving additional diagnostic tools.

The clinical application of a prognostic care decision model is particularly relevant in settings where early and accurate escalation of care is needed. For example, by focusing on these top-5 clinical features or using a decision support tool like ours, even less experienced primary care physicians could assess risk and anticipate tertiary care referrals without advanced diagnostics. Additionally, in emergency settings, these tools could assist in triaging patients to prioritize those needing immediate respiratory support or mechanical ventilation, allowing earlier interventions and more effective resource allocation—crucial for LMICs—potentially reducing morbidity and mortality.

One significant limitation of this study is its reliance on data from a single tertiary hospital (Hacettepe University), which

may limit generalizability. While the dataset includes patients referred from both urban and rural areas, the focus on a tertiary center introduces a selection bias, as most cases represent severe care levels (304/437, 69.6%). This is likely because less severe CAP cases are managed in primary or secondary care, not referred to tertiary centers, limiting the model's applicability in less severe cases. Additionally, the relatively small sample size of 437 patients limits the model's generalizability, as larger datasets are typically needed to optimize ML models and ensure robust performance across diverse populations. Expanding the dataset to include patients from multiple centers, especially primary and secondary care institutions, could improve the model's generalizability and applicability. Lastly, the retrospective nature of the data and the missing time frames of tertiary care unit transfers may not fully capture real-time clinical decision-making or the urgency of care decisions.

In conclusion, this study demonstrates the feasibility of developing an ML-based prognostic decision support tool for childhood pneumonia referral, with an accuracy of 77% to 88%. Incorporating foundational clinical skills for key prognostic predictors with advanced data science methods holds promise for improving pneumonia outcomes by accurately predicting the need for the escalation of care.

Acknowledgments

During the preparation of this work, the authors used OpenAI GPT-4o [41] to restructure sentences for enhanced readability, as they are not native English speakers. After using this tool/service, the authors reviewed and edited the content as needed and took full responsibility for the content of the publication.

Authors' Contributions

OS contributed to the creation of the work plan, interpretation of statistical analysis and machine learning algorithms, coinvestigation of the literature, and writing the revised manuscript. ITA contributed to the building of machine learning algorithms, coinvestigation of the literature, and writing the results and methods. SBC contributed to scanning patients from the hospital electronic health record system and encoding the attributes of the patients' data in the case report form ("Human Encoder-1"). BK contributed to the scanning patients from the hospital electronic health record system and encoding the attributes of the patients' data in the case report form ("Human Encoder-2"). AFD contributed to the building of the machine learning algorithms and optimizing the dataset. MZU contributed to the coding of advanced statistical and machine learning algorithms, and the creation of the clinical decision support system interface. YO contributed to the creation of the work plan, interpretation of statistical analysis, and gathering the team of investigators.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Precision-recall curve (PRC) for the blending model with top 5 features.
[PNG File, 21 KB - xmed_v6i1e57719_app1.png]

Multimedia Appendix 2
Shapley additive explanations (SHAP) values forward selection method.
[PNG File, 386 KB - xmed_v6i1e57719_app2.png]

References

1. Pneumonia in children. World Health Organization. 2022 Nov 11. URL: https://www.who.int/news-room/fact-sheets/detail/pneumonia [accessed 2024-10-01]
2. United Nations Inter-Agency Group for Child Mortality Estimation. Levels and trends in child mortality, report 2023. UNICEF. 2024 Mar 12. URL: https://data.unicef.org/resources/levels-and-trends-in-child-mortality-2024/ [accessed 2024-10-01]
3. Qazi S, Aboubaker S, MacLean R, et al. Ending preventable child deaths from pneumonia and diarrhoea by 2025. development of the integrated Global Action Plan for the Prevention and Control of Pneumonia and Diarrhoea. Arch Dis Child 2015 Feb;100 Suppl 1:S23-S28. [doi: 10.1136/archdischild-2013-305429] [Medline: 25613963]
4. Sazawal S, Black RE, Pneumonia Case Management Trials Group. Effect of pneumonia case management on mortality in neonates, infants, and preschool children: a meta-analysis of community-based trials. Lancet Infect Dis 2003 Sep;3(9):547-556. [doi: 10.1016/s1473-3099(03)00737-0] [Medline: 12954560]
5. Shah SN, Bachur RG, Simel DL, Neuman MI. Does this child have pneumonia?: the rational clinical examination systematic review. JAMA 2017 Aug 1;318(5):462-471. [doi: 10.1001/jama.2017.9039] [Medline: 28763554]
6. World Health Organization. Handbook: IMCI integrated management of childhood illness. World Health Organization. 2005. URL: https://iris.who.int/handle/10665/42939

7. Ferdous F, Ahmed S, Das SK, et al. Pneumonia mortality and healthcare utilization in young children in rural Bangladesh: a prospective verbal autopsy study. Trop Med Health 2018 May 25;46:17. [doi: 10.1186/s41182-018-0099-4] [Medline: 29875615]
8. Shaima SN, Alam T, Bin Shahid A, et al. Prevalence, predictive factors, and outcomes of respiratory failure in children with pneumonia admitted in a developing country. Front Pediatr 2022 May 4;10:841628. [doi: 10.3389/fped.2022.841628] [Medline: 35601439]
9. Sonego M, Pellegrin MC, Becker G, Lazzerini M. Risk factors for mortality from acute lower respiratory infections (ALRI) in children under five years of age in low and middle-income countries: a systematic review and meta-analysis of observational studies. PLoS One 2015 Jan 30;10(1):e0116380. [doi: 10.1371/journal.pone.0116380] [Medline: 25635911]
10. McAllister DA, Liu L, Shi T, et al. Global, regional, and national estimates of pneumonia morbidity and mortality in children younger than 5 years between 2000 and 2015: a systematic analysis. Lancet Glob Health 2019 Jan;7(1):e47-e57. [doi: 10.1016/S2214-109X(18)30408-X] [Medline: 30497986]
11. Chang TH, Liu YC, Lin SR, et al. Clinical characteristics of hospitalized children with community-acquired pneumonia and respiratory infections: Using machine learning approaches to support pathogen prediction at admission. J Microbiol Immunol Infect 2023 Aug;56(4):772-781. [doi: 10.1016/j.jmii.2023.04.011] [Medline: 37246060]
12. Bennett TD, Callahan TJ, Feinstein JA, et al. Data science for child health. J Pediatr 2019 May;208:12-22. [doi: 10.1016/j.jpeds.2018.12.041] [Medline: 30686480]
13. Zhang X, Pérez-Stable EJ, Bourne PE, et al. Big data science: opportunities and challenges to address minority health and health disparities in the 21st century. Ethn Dis 2017 Apr 20;27(2):95-106. [doi: 10.18865/ed.27.2.95] [Medline: 28439179]
14. Sheikh M, Jehan F. Using big data for risk stratification of childhood pneumonia in low-income and middle-income countries (LMICs): challenges and opportunities. EBioMedicine 2021 Dec;74:103740. [doi: 10.1016/j.ebiom.2021.103740] [Medline: 34916165]
15. Liu YC, Cheng HY, Chang TH, et al. Evaluation of the need for intensive care in children with pneumonia: machine learning approach. JMIR Med Inform 2022 Jan 27;10(1):e28934. [doi: 10.2196/28934] [Medline: 35084358]
16. Kanwal K, Khalid SG, Asif M, Zafar F, Qurashi AG. Diagnosis of community-acquired pneumonia in children using photoplethysmography and machine learning-based classifier. Biomed Signal Process Control 2024 Jan;87:105367. [doi: 10.1016/j.bspc.2023.105367]
17. Smith JC, Spann A, McCoy AB, et al. Natural language processing and machine learning to enable clinical decision support for treatment of pediatric pneumonia. AMIA Annu Symp Proc 2020 Jan 25;2020:1130-1139. [Medline: 33936489]
18. Zech JR, Badgeley MA, Liu M, Costa AB, Titano JJ, Oermann EK. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. PLoS Med 2018 Nov 6;15(11):e1002683. [doi: 10.1371/journal.pmed.1002683] [Medline: 30399157]
19. Gera T, Shah D, Garner P, Richardson M, Sachdev HS. Integrated management of childhood illness (IMCI) strategy for children under five. Cochrane Database Syst Rev 2016 Jun 22;2016(6):CD010123. [doi: 10.1002/14651858.CD010123.pub2] [Medline: 27378094]
20. Neyzi O, Bundak R, Gökçay G, et al. Reference values for weight, height, head circumference, and body mass index in Turkish children. J Clin Res Pediatr Endocrinol 2015 Dec;7(4):280-293. [doi: 10.4274/jcrpe.2183] [Medline: 26777039]
21. Mahajan P, Uddin S, Hajati F, Moni MA. Ensemble learning for disease prediction: a review. Healthcare (Basel) 2023 Jun 20;11(12):1808. [doi: 10.3390/healthcare11121808] [Medline: 37372925]
22. Jain S, Williams DJ, Arnold SR, et al. Community-acquired pneumonia requiring hospitalization among U.S. children. N Engl J Med 2015 Feb 26;372(9):835-845. [doi: 10.1056/NEJMoa1405870] [Medline: 25714161]
23. Shao JH, Yu KH, Chen SH. COVID-19-related disruptions in implementation of a randomized control trial: an autoethnographic report. Appl Nurs Res 2023 Aug;72:151698. [doi: 10.1016/j.apnr.2023.151698] [Medline: 37423680]
24. Sohrabi C, Mathew G, Franchi T, et al. Impact of the coronavirus (COVID-19) pandemic on scientific research and implications for clinical academic training - a review. Int J Surg 2021 Feb;86:57-63. [doi: 10.1016/j.ijsu.2020.12.008] [Medline: 33444873]
25. Kuitunen I, Artama M, Mäkelä L, Backman K, Heiskanen-Kosma T, Renko M. Effect of social distancing due to the COVID-19 pandemic on the incidence of viral respiratory tract infections in children in Finland during early 2020. Pediatr Infect Dis J 2020 Dec;39(12):e423-e427. [doi: 10.1097/INF.0000000000002845] [Medline: 32773660]
26. Chen M, Zhou Y, Jin S, et al. Changing clinical characteristics of pediatric inpatients with pneumonia during COVID-19 pandamic: a retrospective study. Ital J Pediatr 2024 Apr 23;50(1):84. [doi: 10.1186/s13052-024-01651-8] [Medline: 38650007]
27. Huang C. Pediatric non-COVID-19 community-acquired pneumonia in COVID-19 pandemic. Int J Gen Med 2021 Oct 27;14:7165-7171. [doi: 10.2147/IJGM.S333751] [Medline: 34737611]
28. Lastrucci V, Bonaccorsi G, Forni S, et al. The indirect impact of COVID-19 large-scale containment measures on the incidence of community-acquired pneumonia in older people: a region-wide population-based study in Tuscany, Italy. Int J Infect Dis 2021 Aug;109:182-188. [doi: 10.1016/j.ijid.2021.06.058] [Medline: 34216731]
29. Latif S, Usman M, Manzoor S, et al. Leveraging data science to combat COVID-19: a comprehensive review. IEEE Trans Artif Intell 2020 Sep 2;1(1):85-103. [doi: 10.1109/TAI.2020.3020521] [Medline: 37982070]

30. Hu S, Wang X, Ma Y, Cheng H. Global research trends in pediatric COVID-19: a bibliometric analysis. Front Public Health 2022 Feb 16;10:798005. [doi: 10.3389/fpubh.2022.798005] [Medline: 35252087]
31. Chumbita M, Cillóniz C, Puerta-Alcalde P, et al. Can artificial intelligence improve the management of pneumonia. J Clin Med 2020 Jan 17;9(1):248. [doi: 10.3390/jcm9010248] [Medline: 31963480]
32. Bradley JS, Byington CL, Shah SS, et al. The management of community-acquired pneumonia in infants and children older than 3 months of age: clinical practice guidelines by the Pediatric Infectious Diseases Society and the Infectious Diseases Society of America. Clin Infect Dis 2011 Oct;53(7):e25-e76. [doi: 10.1093/cid/cir531] [Medline: 21880587]
33. Harris M, Clark J, Coote N, et al. British Thoracic Society guidelines for the management of community acquired pneumonia in children: update 2011. Thorax 2011 Oct;66 Suppl 2:ii1-i23. [doi: 10.1136/thoraxjnl-2011-200598] [Medline: 21903691]
34. Dean P, Florin TA. Factors associated with pneumonia severity in children: a systematic review. J Pediatric Infect Dis Soc 2018 Dec 3;7(4):323-334. [doi: 10.1093/jpids/piy046] [Medline: 29850828]
35. Araya S, Lovera D, Zarate C, et al. Application of a prognostic scale to estimate the mortality of children hospitalized with community-acquired pneumonia. Pediatr Infect Dis J 2016 Apr;35(4):369-373. [doi: 10.1097/INF.0000000000001018] [Medline: 26629871]
36. Williams DJ, Zhu Y, Grijalva CG, et al. Predicting severe pneumonia outcomes in children. Pediatrics 2016 Oct;138(4):e20161019. [doi: 10.1542/peds.2016-1019] [Medline: 27688362]
37. Zeng M, Zou B, Wei F, Liu X, Wang L. Effective prediction of three common diseases by combining SMOTE with Tomek links technique for imbalanced medical data. Presented at: 2016 IEEE International Conference of Online Analysis and Computing Science (ICOACS); May 28-29, 2016; Chongqing, China p. 225-228. [doi: 10.1109/ICOACS.2016.7563084]
38. Liu R, Greenstein JL, Fackler JC, Bergmann J, Bembea MM, Winslow RL. Prediction of impending septic shock in children with sepsis. Crit Care Explor 2021 Jun 15;3(6):e0442. [doi: 10.1097/CCE.0000000000000442] [Medline: 34151278]
39. Akhtar F, Li J, Pei Y, Xu Y, Rajput A, Wang Q. Optimal features subset selection for large for gestational age classification using GridSearch based recursive feature elimination with cross-validation scheme. In: Hung J, Yen N, Chang JW, editors. Frontier Computing: Theory, Technologies and Applications (FC 2019). Lecture Notes in Electrical Engineering, vol 551: Springer; 2020:63-71. [doi: 10.1007/978-981-15-3250-4_8]
40. Man X, Chan EP. The best way to select features? comparing MDA, LIME, and SHAP. J Financ Data Sci Winter 2021;3(1):127-139. [doi: 10.3905/jfds.2020.1.047]
41. GPT-4o. OpenAI. URL: https://platform.openai.com/docs/models/gpt-4o [accessed 2025-02-12]

Abbreviations

AUC-ROC: area under the receiver operator characteristic curve
CAP: community-acquired pneumonia
IMCI: Integrated Management of Childhood Illness
LightGBM: light gradient boosting machine
LMIC: low- and middle-income country
ML: machine learning
RFECV: recursive feature elimination with cross-validation
SHAP: Shapley additive explanations
SMOTE: Synthetic Minority Oversampling Technique
WHO: World Health Organization
XGBoost: extreme gradient boosting

Edited by S Amal; submitted 24.02.24; peer-reviewed by Anonymous, C Rogerson; revised version received 19.12.24; accepted 08.01.25; published 04.03.25.

Please cite as:
Serin O, Akbasli IT, Cetin SB, Koseoglu B, Deveci AF, Ugur MZ, Ozsurekci Y
Predicting Escalation of Care for Childhood Pneumonia Using Machine Learning: Retrospective Analysis and Model Development
JMIRx Med 2025;6:e57719
URL: https://xmed.jmir.org/2025/1/e57719
doi:10.2196/57719

Financial Feasibility of Developing Sustained-Release Incrementally Modified Drugs in Thailand's Pharmaceutical Industry: Mixed Methods Study

Manthana Laichapis[1], PharmD; Rungpetch Sakulbumrungsil[1], PhD; Khunjira Udomaksorn[2], PhD; Nusaraporn Kessomboon[3], PhD; Osot Nerapusee[1], PhD; Charkkrit Hongthong[3], MSc; Sitanun Poonpolsub[4], PharmD

[1]Department of Social and Administrative Pharmacy, Faculty of Pharmaceutical Sciences, Chulalongkorn University, 254 Phayathai Road, Pathum Wan, Bangkok, Thailand

[2]Department of Social and Administrative Pharmacy, Faculty of Pharmaceutical Sciences, Prince of Songkla University, Songkla, Thailand

[3]Department of Social and Administrative Pharmacy, Faculty of Pharmaceutical Sciences, Khon Kaen University, Khon Kaen, Thailand

[4]Food and Drug Administration Thailand, Nonthaburi, Thailand

Corresponding Author:
Rungpetch Sakulbumrungsil, PhD
Department of Social and Administrative Pharmacy, Faculty of Pharmaceutical Sciences, Chulalongkorn University, 254 Phayathai Road, Pathum Wan, Bangkok, Thailand

Related Articles:

Companion article: https://www.medrxiv.org/content/10.1101/2024.07.29.24311184v1

Companion article: https://med.jmirx.org/2025/1/e78090

Companion article: https://med.jmirx.org/2025/1/e77627

Companion article: https://med.jmirx.org/2025/1/e77623

Abstract

Background: Thailand's pharmaceutical industry is prioritizing innovation and self-reliance through the development of incrementally modified drugs (IMDs), particularly sustained-release dosage forms. However, the financial feasibility of IMD development remains underexplored.

Objective: This study evaluates the financial feasibility of developing sustained-release IMDs in Thailand, focusing on costs, timelines, and investment requirements to inform strategic decision-making.

Methods: A mixed methods approach was used, combining literature reviews, expert interviews, and financial modeling. Two scenarios were analyzed: (1) only development (phase I) and (2) full clinical trials (phase I to III). Sensitivity analysis was used to assess the impact of key variables on financial feasibility.

Results: The research and development (R&D) process for sustained-release IMDs takes 7 years for phase I–only development, costing US $1.46 - 3.09 million, and 11 years for full clinical trials, costing US $18.60 - 20.23 million. Process validation batches accounted for 60% of costs in phase I–only scenarios, while clinical trials represented 70% of costs in full clinical trial scenarios. The annual income required for a 5-year payback period ranged from US $0.20 - 1.80 million (phase I only) to US $3.01 - 27.11 million (full trials). Shorter R&D durations and longer payback periods substantially improved feasibility.

Conclusions: Developing sustained-release IMDs in Thailand involves substantial costs and extended timelines but offers a lower-risk alternative to new chemical entities. Strategic investments, efficient R&D processes, and supportive policies are essential to enhance feasibility and alignment with national goals of innovation and self-reliance.

(JMIRx Med 2025;6:e65978) doi:10.2196/65978

KEYWORDS

financial; economics; R&D; research and development; surveys; interviews; costs; revenue; policies; drugs; pharmaceuticals

Introduction

The Thai pharmaceutical industry is undergoing significant transformation in alignment with Thailand's Pharmaceutical Development Action Plan (2023 - 2027), which builds upon the foundation of earlier policies, including the National Strategic Master Plan (2018 - 2037). These initiatives collectively emphasize enhancing the national drug system, fostering domestic pharmaceutical manufacturing capabilities, and achieving self-reliance through sustainable development [1-3]. The current action plan (2023 - 2027) prioritizes accelerating the industry's capabilities in research, development, and production of vaccines, drugs, herbal products, and biologics, while promoting local pharmaceutical industries to reduce import dependency and increase export potential. By focusing on innovation, technological advancement, and strategic investments, this plan ensures the industry's alignment with national health care priorities and global market demands, driving growth and competitiveness in the pharmaceutical sector.

Currently, Thailand's pharmaceutical manufacturing industry is predominantly focused on the production of generic drugs, with an average of 540 generic drug approvals annually, including approximately 35 new ones [3]. However, the development of new chemical entities (NCEs) remains limited due to challenges such as insufficient investment, lack of advanced technology, and a shortage of specialized talent. Given these constraints, a more feasible approach for Thailand's pharmaceutical sector lies in the development of incrementally modified drugs (IMDs). IMDs involve enhancing existing drugs through modifications in delivery systems, indications, combinations, administration routes, dosage forms, and strengths, offering a pathway to sustainable self-reliance while reducing costs and risks associated with NCE development [4].

Globally, IMDs have gained traction in high-income countries, with high listing and reimbursement rates, demonstrating their potential to improve patient outcomes and health care efficiency [5]. In Thailand, focusing on IMDs aligns with the country's strategic goals of fostering innovation, reducing reliance on imported pharmaceuticals, and enhancing the competitiveness among local manufacturers. By leveraging advanced technology platforms, the development of IMDs can provide a viable pathway for the Thai pharmaceutical industry to achieve greater self-sufficiency and contribute to the broader health care system.

This study aims to examine the financial feasibility of developing IMD dosage forms within Thailand's pharmaceutical manufacturing industry. By evaluating the economic viability and potential return on investment of IMDs, this study will provide evidence-based insights to support decision-making and guide strategic investments in the sector. The findings may provide a foundation for policy makers and stakeholders in formulating targeted strategies to promote innovation in IMDs, enhance domestic pharmaceutical capabilities, and reduce reliance on imported drugs. Ultimately, the study may also contribute to strengthening Thailand's pharmaceutical sector, ensuring its alignment with national development goals and global health care trends.

Methods

Study Design

This study used a mixed methods approach, combining qualitative and quantitative components. Given the lack of publicly available data on IMD development, this approach was necessary to triangulate data from multiple sources, ensuring robustness and reliability. The qualitative component included a literature review, surveys, and expert interviews, while the quantitative component focused on financial modeling and analysis.

Data Collection

Literature Review

A comprehensive review of existing IMD dosage forms, manufacturing processes, cost structures, regulatory requirements, and market trends was conducted using PubMed, Scopus, and industry reports. This review served as input for the development of the financial model and interview guide.

Survey

A survey was designed to estimate costs associated with IMD development. The cost structures were adapted from a prior study on the impact of the Thai–European Union (EU) free trade agreement (FTA) on the pharmaceutical supply chain in Thailand [6]. Cost collection forms were sent to five IMD experts for feedback, refinement, and validation, after which cost estimates were provided.

Interviews

Snowball sampling was used to identify participants due to the specialized nature of IMD development and the limited number of manufacturers in this field. This approach allowed the research team to access experts with relevant knowledge and experience in IMD development.

Semistructured interviews were conducted with 15 experts, including company owners, industry leaders, policy makers, and researchers. Interviews continued until data saturation was achieved, with no new themes emerging.

Interviews were conducted online and recorded with participants' consent. The key interview questions were focused on costs associated with research and development (R&D), manufacturing technology, and clinical and nonclinical studies. Data saturation was achieved when no new themes emerged. To ensure transparency and reproducibility, detailed descriptions of the interview guide and survey questions are provided in Multimedia Appendix 1.

Ethical Considerations

The study received ethical approval from the Research Ethics Review Committee for Research Involving Human Subjects, Health Science Group at Chulalongkorn University, Thailand (COA No. 176/2564). We confirm that participation in the online Zoom sessions was entirely voluntary. Participants were informed in advance about the purpose and format of the sessions, and they had the right to decline participation without any consequences. Choosing to join the Zoom session was

considered as implied consent by action, in alignment with ethical practices for minimal-risk research. To protect participant confidentiality, no personally identifiable information was recorded during the sessions. All data were de-identified prior to analysis. Participants received approximately $30 USD as compensation for their time and contribution, ensuring transparency and fairness throughout the research process.

We confirm that participation in the online Zoom sessions was entirely voluntary. Participants were informed in advance about the purpose and format of the sessions, and they had the right to decline participation without any consequences. Choosing to join the Zoom session was considered as **implied consent by action**, in alignment with ethical practices for minimal-risk research.

Data Analysis

Financial Model Development

Two financial model scenarios were developed to assess IMD feasibility, focusing on IMD types and cost estimation. Sustained-release formulations, which are the most preferred type of IMDs by the domestic pharmaceutical industry, were selected based on results from a prior feasibility study [7]. The cost estimation was adapted from the Thai-EU FTA study [6], covering sourcing, R&D (ie, laboratory scale, pilot batch, and stability studies), nonclinical and clinical trials, and registration and process validation [8,9]. Costs were estimated under two regulatory scenarios: (1) conducting only phase I clinical trials and (2) conducting full clinical trials [10].

Sensitivity Analysis

A sensitivity analysis was conducted to evaluate the impact of variations in key variables including cost, duration, and payback period on the financial feasibility of IMD development. This analysis provided insights into the robustness of the financial models under varying assumptions.

Results

Overview

A prediction market analysis by Hongthong et al [11] on the feasibility of IMD development by the domestic pharmaceutical industry identified sustained-release dosage forms as the most preferred option, which guided the financial feasibility analysis in this study [7,12]. The assumptions and input data used for this analysis are detailed in Table 1.

Table . Input data and assumptions for the financial model and financial feasibility study.

Variables	Assumptions	Source of data
Cost of sales	25% of revenue	Jiang et al [13]
Operational expense	40% of revenue	Jiang et al [13] and interviews
Discounted rate	Discount rate of 3%	Haacker et al [12]
Interest rate	Interest rate for business is 3%	Interviews
Tax rate	Corporate tax rate is 20%	IDRG Consultancy Services [14] and interviews
Expected payback period	Payback period that investors could accept is 5 - 10 years	Interviews

Financial Feasibility Analysis Model

To assess the financial viability of investing in the development of sustained-release dosage forms, a financial feasibility analysis model was developed. This model calculates the payback period and market growth rate based on two primary components (ie, cost and revenue components). The cost component estimates the expenses associated with R&D of the new dosage form, while the revenue component forecasts the income required to achieve a return on investment within a specified payback period. The model's flexibility allows for adjustments in key variables such as the payback period and market growth rate, enabling stakeholders to make informed strategic decisions regarding pharmaceutical R&D investments.

Cost Analysis for Sustained-Release Dosage Form Development

Table 2 presents the cost analysis for two development scenarios. In scenario 1, which involves only phase I clinical studies, the R&D process for new sustained-release IMD formulations was estimated to take 7 years, with development costs ranging from US $1.46 to 3.09 million. Approximately 60% of the total costs were allocated to process validation batches, a critical step requiring three consecutive production batches. This phase represents a significant capital investment, with varying costs depending on production complexity. In scenario 2, which includes full clinical trials from phase I to phase III, the development duration extended to 11 years, with fixed costs ranging from US $18.60 to 20.23 million. In this case, 70% of the total R&D budget was dedicated to clinical studies, which are essential for demonstrating the efficacy and safety of new drugs.

The sensitivity analysis presented in Table 3 evaluates the financial implications under different scenarios. For scenario 1, which involves only phase I studies with R&D costs of US $1.46 million, the annual income required to recover the invested capital—assuming a 5-year payback period—ranges from US $0.20 to 1.80 million. In contrast, for scenario 2, which includes full clinical trials, the required annual income increases substantially, ranging from US $3.01 to 27.11 million.

Table . The process and cost of developing IMDs[a] in a sustained-release form by the domestic pharmaceutical industry.

Process	Details [6]	Scenario 1 (US $, in millions)	Scenario 2 (US $, in millions)	Information source
Sourcing	Local manufacturers choose reference IMDs based on marketing, user needs, sales, patents, and suitability.	0.02	0.02	This study
R&D[b] laboratory scale	The R&D department conducts laboratory-scale studies to develop suitable formulations for sustained-release drugs, including analytical method development and determination of finished product specifications (FPS).	0.06 - 0.15	0.06 - 0.15	This study
Pilot scale	After successful drug R&D, pilot batch production begins, followed by stability studies to determine shelf-life specifications. Results are reported to the FDA[c].	0.27 - 0.66	0.27 - 0.66	Sertkaya et al [15] and this study
Clinical study				
Phase I	Samples from the pilot batch production will be sent to study the effect of food on bioefficacy through a bioequivalence study and evaluating the effect of alcohol on dose dumping.	0.29	0.29	Thai FDA [10], National Institute of Health [16], DiMasi et al [17], and this study
Phase II	In case the pharmacokinetics of a new drug are clinically significantly different from the reference drug, phase II and III studies of the new drug may be necessary.	—[e]	4.29	Thai FDA [10] and this study
Phase III	In case the pharmacokinetics of a new drug are clinically significantly different from the reference drug, phase II and III studies of the new drug may be necessary.	—	12.86	Thai FDA [10] and this study
Registration	Registration of new drug formulas follows the ASEAN[d] harmonization criteria. Application documents included administration data, product information, quality, safety, and efficacy parts. The nonclinical and clinical study data can refer to recommendations and guidelines.	0.003	0.003	This study
Process validation batch	After obtaining FDA registration, drugs can be produced in commercial batches. Process validation and inspection results are submitted for permission to continue production and distribution.	0.81 - 1.97	0.81 - 1.97	This study
Total cost	—	1.46 - 3.09	18.60 - 20.23	—

[a]IMD: incrementally modified drug.

[b]R&D: research and development.

[c]FDA: Food and Drug Administration.

[d]ASEAN: Association of Southeast Asian Nations.

[e]Not applicable.

Table . Expected annual revenue after product launch of both scenarios (US $ [in millions]/year).

Expected payback	Annual revenue (US $ [in millions]/year)									
	Year 1	Year 2	Year 3	Year 4	Year 5	Year 6	Year 7	Year 8	Year 9	Year 10
Scenario 1[a]										
5-year period	0.20	0.40	0.80	1.20	1.80	—[c]	—	—	—	—
10-year period	0.04	0.08	0.17	0.25	0.38	0.49	0.61	0.77	0.96	1.20
Scenario 2[b]										
5-year period	3.01	6.03	12.05	18.08	27.11	—	—	—	—	—
10-year period	0.63	1.25	2.50	3.75	5.63	7.32	9.15	11.44	14.30	17.87

[a]The parameters of the base case analysis for scenario 1 are that the duration of research and development is 7 years, and the research and development cost is US $1.46 million.

[b]The parameters of the base case analysis for scenario 2 are that the duration of research and development is 11 years, and the research and development cost is US $18.60 million.

[c]Not applicable.

The duration of the R&D process substantially impacts financial expenditures; shorter R&D periods can reduce costs and enhance project feasibility. High-risk activities such as complex formulation and analytical method development often involve a higher likelihood of failure and require advanced clinical studies, necessitating larger capital investments. Therefore, well-planned R&D processes can substantially reduce costs and improve investment returns.

The analysis also examines the impact of extending the payback period to 10 years, which lowers the capitalization point and reduces the annual income required. This consideration is particularly relevant for drugs targeting chronic diseases, which typically have longer market life cycles. Additionally, factors such as annual sales growth rates upon launch, the competitive landscape, and government regulations critically influence the financial feasibility of developing sustained-release IMDs [15].

Discussion

Principal Findings

This study systematically explores the financial viability and strategic implications of developing IMDs, with a focus on sustained-release dosage forms. Our analysis highlights the financial and investment requirements for launching IMDs into the market, particularly in comparison to generic drugs.

Developing IMDs, particularly as sustained-release formulations, is substantially more resource-intensive and time-consuming than producing generic drugs. The extended timelines and higher costs are primarily attributed to the complexities of modifying and validating existing drugs, which necessitate extensive clinical testing. In contrast, Liangrokapart et al [6], in a study on the impact of the Thai-EU FTA concerning intellectual property rights on the pharmaceutical supply chain in Thailand, suggested that generic drug development typically required 25 to 46 months, with considerably lower R&D costs ranging from US $0.19 to 1.13 million [6].

A key challenge in this analysis is the absence of specific active ingredient data, which complicates accurate forecasting of market growth and sales revenue. Despite these limitations, the chosen methodology effectively captures the financial intricacies of IMD development, providing robust insights into the associated costs and investment requirements.

Compared to the development costs of NCEs reported in previous studies—including an analysis by Sertkaya et al [15] on drug development costs in the United States (2000 - 2018) and the study by DiMasi et al [18] on the price of innovation and research on R&D costs and returns by therapeutic category—the costs for IMDs are substantially lower. This is primarily because IMDs do not incur discovery and preclinical expenses. Additionally, IMDs have lower failure rates than NCEs, suggesting a potentially lower-risk investment profile. This aligns with findings from a study on IMDs under the USFDA 505(b)(2) NDA pathway, which reported clinical trial completion times of 12 - 24 months and development costs of US $2 to 10 million, closely mirroring the outcomes of this study [4].

The findings from both scenarios underscore that IMD development entails higher costs and longer timelines compared

to generic drugs. These challenges stem from the need to develop new formulations and conduct comprehensive clinical studies. However, shorter R&D periods can substantially reduce costs and enhance project feasibility, emphasizing the importance of efficient R&D planning.

The early stage and inexperience of IMD development within the domestic industry may result in longer timelines and elevated costs. Limited domestic expertise, coupled with the complexities of clinical trials and regulatory processes, poses additional challenges. The high investment required for IMD development necessitates a strong focus on market feasibility and sales potential, particularly in a competitive landscape dominated by generic drugs.

While this study offers valuable insights into the financial feasibility of IMD development in Thailand, several limitations must be acknowledged. First, cost estimates were derived from expert feedback and prior studies, which may not fully capture the variability inherent in real-world manufacturing processes. Finally, the findings may be context specific and not directly applicable to other types of IMDs or pharmaceutical markets.

For the future direction of this research, IMDs represent an incremental innovation that can be developed in various forms, including stand-alone and combination products. Therefore, further studies are needed to assess the feasibility of developing different types of IMDs to enhance patient health outcomes and quality of life.

Additionally, the regulatory process and guidelines play a crucial role in IMD development, making it necessary to study the impact of regulatory changes on IMDs. Furthermore, the pricing and reimbursement mechanisms for IMDs remain unclear for the local pharmaceutical industry, highlighting the need for further exploration of this topic.

Conclusions

This study provides essential insights into the financial aspects of developing sustained-release IMDs in Thailand, highlighting the extensive resources and strategic planning required. These findings underscore the complexity of predicting financial outcomes due to the variability in active ingredients and market dynamics. Although the development of IMDs involves substantial investment and extended timelines, understanding these financial and operational dimensions is crucial for successful drug development. Future research should further investigate the full cost spectrum of various types of IMD approaches to enhance the financial predictability and success of these studies.

Acknowledgments

The authors would like to express our sincere gratitude to the Thai Pharmaceutical Manufacturers Association (TPMA) for their excellent collaboration, which allowed us to conduct interviews and gather valuable data for the assessment. Our deepest appreciation goes to all the participants who generously gave their time and contributed to this study. Their active participation and valuable insights greatly enriched our research.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Survey instrument and interview guide for the financial feasibility study of sustained-release incrementally modified drugs in Thailand.
[DOCX File, 17 KB - xmed_v6i1e65978_app1.docx]

References

1. The national strategy of Thailand (2018 - 2037) [Thai]. : Secretariat of the National Strategy Committee; 2016 URL: https://dl.parliament.go.th/handle/20.500.13072/573515 [accessed 2024-08-18]
2. Jitruknatee A, Martro J, Tosanguan K, Doangjai Y, Theantawee W. National drug policies in Thailand: evolution and lessons for the future. J Health Sci Thai 2020 Jan 27;29:S3-S14 [FREE Full text]
3. Committee on Thai Drug System. Thai drug system 2020. : Health System Research Institute (HSRI); 2020 URL: https://www.hsri.or.th/printed-matter/412 [accessed 2024-08-18]
4. Ilenwabor RW. Incrementally modifying drugs via changing route of administration. J Pharm Dev Ind Pharm 2022;4(2):16-29 [FREE Full text]
5. Ha D, Choi Y, Kim DU, Chung KH, Lee EK. A comparative analysis of the impact of a positive list system on new chemical entity drugs and incrementally modified drugs in South Korea. Clin Ther 2011 Jul;33(7):926-932. [doi: 10.1016/j.clinthera.2011.05.089] [Medline: 21715008]
6. Liangrokapart J, Kessomboon N, Sakulbumrungsil R, Akaleephan C, Poonpolsub S, Saerekul P. Impact of Thai-EU free trade agreement (FTA) concerning intellectual property rights on the pharmaceutical supply chain in Thailand. 2013.
7. Hongthong C, Sakulbumrungsil R, Udomaksorn K, et al. Feasibility study of sustained release dosage forms for incrementally modified drug by domestics pharmaceutical industry in Thailand. F1000Res 2023;12:1513. [doi: 10.12688/f1000research.142745.1] [Medline: 39512910]

8. Abraham S, Tekwani K, Goyal K, Sujatha C. A study on the role of cost accounting technique in manufacturing industries. 2022 URL: https://www.researchgate.net/publication/359847680_A_STUDY_ON_THE_ROLE_OF_COST_ACCOUNTING_TECHNIQUE_IN_MANUFACTURING_INDUSTRIES [accessed 2025-06-30]
9. Chotayakul S, Punyangarm V. Manufacturing cost estimation design to estimate the production cost of new products. J Adv Develop Eng Sci 2023;10(29):15-28 [FREE Full text]
10. Guideline for nonclinical and clinical evaluations of new drug products from previously approved drug substances. Thai Food and Drug Administration. 2019. URL: https://en.fda.moph.go.th/entrepreneurs-medicines/category/how-to-apply-for-drug-approval/ [accessed 2025-06-19]
11. Hongthong C, Sakulbumrungsil R, Kessomboon N, et al. Dosage form selection using price function of prediction market. F1000Res 2025;14:197. [doi: 10.12688/f1000research.161732.1]
12. Haacker M, Hallett TB, Atun R. On discount rates for economic evaluations in global health. Health Policy Plan 2020 Feb 1;35(1):107-114. [doi: 10.1093/heapol/czz127] [Medline: 31625564]
13. Jiang J, Kong J, Grogan J. How did the public US drugmakers' sales, expenses and profit change over time? USC Leonard D Schaeffer Institute for Public Policy & Government Service. 2021. URL: https://healthpolicy.usc.edu/evidence-base/how-did-the-public-u-s-drugmakers-sales-expenses-and-profits-change-over-time [accessed 2024-08-18]
14. Department of Industries Ministry of Economic Affairs Royal Government of Bhutan. Feasibility analysis of drug formulation. : IDRG Consultancy Services; 2009. URL: https://www.moice.gov.bt/wp-content/uploads/2023/03/hospitaL-PART-2.pdf [accessed 2025-06-19]
15. Sertkaya A, Beleche T, Jessup A, Sommers BD. Costs of drug development and research and development intensity in the US, 2000-2018. JAMA Netw Open 2024 Jun 3;7(6):e2415445. [doi: 10.1001/jamanetworkopen.2024.15445] [Medline: 38941099]
16. What are clinical trials and studies? National Institute on Aging. 2023. URL: https://www.nia.nih.gov/health/clinical-trials-and-studies/what-are-clinical-trials-and-studies [accessed 2025-06-15]
17. DiMasi JA, Hansen RW, Grabowski HG. The price of innovation: new estimates of drug development costs. J Health Econ 2003 Mar;22(2):151-185. [doi: 10.1016/S0167-6296(02)00126-1] [Medline: 12606142]
18. DiMasi JA, Grabowski HG, Vernon J. R&D costs and returns by therapeutic category. Drug Information J 2004 Jul;38(3):211-223. [doi: 10.1177/009286150403800301]

Abbreviations

EU: European Union
FTA: free trade agreement
IMD: incrementally modified drug
NCE: new chemical entity
R&D: research and development

Edited by A Grover; submitted 30.08.24; peer-reviewed by E Shkarupeta, P Luksameesate; revised version received 27.02.25; accepted 04.05.25; published 01.07.25.

Please cite as:
Laichapis M, Sakulbumrungsil R, Udomaksorn K, Kessomboon N, Nerapusee O, Hongthong C, Poonpolsub S
Financial Feasibility of Developing Sustained-Release Incrementally Modified Drugs in Thailand's Pharmaceutical Industry: Mixed Methods Study
JMIRx Med 2025;6:e65978
URL: https://xmed.jmir.org/2025/1/e65978
doi:10.2196/65978

Determinants of Periodic Health Examination Uptake: Insights From a Jordanian Cross-Sectional Study

Abdul Aziz Tayoun, MPH

School of Medicine, Department of Family and Community Medicine, Jordan University, Qween Rania Street, Amman, Jordan

Corresponding Author:
Abdul Aziz Tayoun, MPH
School of Medicine, Department of Family and Community Medicine, Jordan University, Qween Rania Street, Amman, Jordan

Related Articles:

Companion article: https://www.medrxiv.org/content/10.1101/2024.02.03.24302286v1

Companion article: https://med.jmirx.org/2025/1/e71529

Companion article: https://med.jmirx.org/2025/1/e71531

Companion article: https://med.jmirx.org/2025/1/e71528

Abstract

Background: Routine periodic health examinations (PHEs) for adults who are asymptomatic are included in clinical preventive services. They aim to prevent morbidity and mortality by identifying modifiable risk factors and early signs of treatable diseases. PHEs are a standard procedure in primary health care worldwide, including in Jordan. The country is undergoing an epidemiological transition toward noncommunicable diseases, which are the leading causes of morbidity and mortality. The prevalence of smoking is among the highest in the world, with escalating rates of obesity and physical inactivity. Notably, hypertension and diabetes are the most prevalent diseases.

Objective: This study aims to determine the extent to which individuals in Jordan participate in PHEs and to evaluate the various factors related to sociodemographics, health, knowledge, and behavior that influence this participation.

Methods: This study used a cross-sectional design and includes 362 participants 18 years or older residing in Jordan. A convenience sampling method was used, and data were collected through a hybrid web-based and face-to-face questionnaire. The analysis involved the application of logistic regression through SPSS to investigate the relationship between various influencing factors and the uptake of PHEs.

Results: Our study indicated that only 98 of the 362 (27.1%, 95% CI 22.8%-31.9%) participants underwent PHEs within the last 2 years. Noteworthy predictors of PHE uptake among Jordanians included recent visits to a primary health care facility within the previous year (adjusted odds ratio [AOR] 4.32, 95% CI 2.40 - 7.76; *P*<.001), monthly income (*P*=.02; individuals with a monthly income of 1500 - 2000 JD displayed more than five times the odds of undertaking PHEs than those with a monthly income <500 JD; AOR 5.74, 95% CI 1.32 - 24.90; *P*=.02; those with a monthly income of more than 2000 JD exhibited even higher odds; AOR 9.81, 95% CI 1.73 - 55.55; *P*=.02; a currency exchange rate of 1 JD=US $1.43 is applicable), and knowledge levels regarding PHEs and preventive health measures (AOR 1.23, 95% CI 1.03 - 1.47; *P*=.007). These variables emerged as the strongest predictors in our analysis, shedding light on key factors influencing PHE uptake in the population. Contrary to other research, our study did not find any statistically significant association between gender (*P*=.33), smoking status (*P*=.76), marital status (*P*=.52), health status self-evaluation (*P*=.18), seasonal influenza vaccination (*P*=.07), combined health behavior factors (*P*=.34), and BMI (*P*=.76) and PHE uptake.

Conclusions: PHE uptake is notably low in Jordan. Critical determinants of this uptake include recent visits to a primary health care facility within the previous year, monthly income, and knowledge levels regarding PHEs and preventive health services. To enhance PHE uptake, there is a critical need to integrate PHEs with primary health care services, increase awareness about PHEs, and offer free preventive services, particularly for those at high risk.

(JMIRx Med 2025;6:e57597) doi:10.2196/57597

KEYWORDS

periodic health examination; PHE; preventive health services; routine health checkups; Jordan; cross-sectional study

Introduction

Background

Routine periodic health examinations (PHEs) for adults who are asymptomatic are integral to primary health care practice. These examinations involve clinical preventive services administered by primary health care clinicians to individuals without signs or symptoms of illness, constituting a routine health care process. The goal of these examinations is to prevent morbidity and mortality proactively, this is achieved by identifying modifiable risk factors and detecting early signs of treatable diseases [1].

The health belief model (HBM) was conceptualized to elucidate why individuals are reluctant to engage in disease prevention programs and health checkups. As a crucial predictive framework, the HBM aids in understanding various health-related behaviors, including smoking, exercise, patient roles, and use of medical services [2].

Integrating with the HBM, health beliefs are defined as personal convictions associated with perceiving and managing specific diseases. These beliefs encompass key elements: perceived sensitivity, perceived severity, perceived benefit, perceived barrier, and cue to action [3].

Literature Review

A systematic review recently published in the *Canadian Family Physician Journal* aimed to assess the reasons for visits to primary health care clinics. Clinicians participating in the review identified routine health maintenance as the third most prevalent reason for individuals seeking consultations with primary health care physicians. This ranking positioned routine health maintenance after upper respiratory tract infections and hypertension, highlighting the significant role of primary health care practitioners in motivating individuals to engage with preventive health services [4].

A study conducted among undergraduate students in a Nigerian health science college found that 91.2% of participants demonstrated awareness of PHEs. However, the actual participation in PHEs was notably low at 28.4%. The primary obstacles to uptake were identified as insufficient time, religious considerations, duration of education, perceived susceptibility to diseases, financial constraints, apprehension about the results, and a general lack of interest [5].

A nationwide study in Saudi Arabia revealed that 22.9% of participants 15 years or older had undergone a PHE in the preceding 2 years. The probability of receiving a PHE during this period exhibited positive correlations with various factors—including age; educational attainment; marital status; regular consumption of five servings of fruits and vegetables daily; and diagnoses such as prediabetes, diabetes, or hypercholesterolemia—visit to a health care setting within the last 2 years due to illness or injury [6].

Rationale and Significance of the Study

Jordan, classified as an upper middle–income country, spans an area of 89,318 square kilometers and is divided into four provinces and 12 governorates. The population has grown substantially, increasing from 5.4 million in 2003 to over 11.5 million in 2023. This demographic shift can be attributed mainly to the influx of refugees and a relatively high birth rate [7,8].

The country is undergoing a notable epidemiological transition characterized by a rising prevalence of noncommunicable diseases (NCDs). These diseases are responsible for approximately 78% of deaths, establishing themselves as the primary cause of mortality and morbidity among the Jordanian population. Key risk factors contributing to the burden of NCDs include tobacco use, with a prevalence of about 50% (including e-cigarettes and shisha). One-quarter of the population reports insufficient physical activity and approximately 60% are classified as overweight or obese. Additionally, 22% of the population has hypertension, 14% has diabetes, and about 18% has depression [9].

Goals of This Study

This profile underscores a pressing concern regarding the country's high risk of NCDs. There is a need for evidence-based preventive health measures to curb the progression of NCDs and their associated risk factors. If conducted according to evidence-based guidelines, PHEs can effectively control communicable diseases and NCDs. Recognizing the urgency of the situation, gathering data on the uptake rate of PHEs, and identifying the factors influencing this uptake is imperative. The absence of previous studies on the uptake of PHEs in Jordan underscores the necessity for comprehensive research. Our study aims to estimate the uptake of PHEs among Jordanians while concurrently investigating various sociodemographic, health status, knowledge, and behavioral factors that play a role in influencing this uptake. The findings from this research will not only contribute valuable insights into the current scenario but also guide educational and promotional activities to encourage citizens to use preventive health services. In doing so, we strive to fill a crucial gap in existing knowledge and provide a foundation for evidence-based strategies to enhance public health in the country.

Methods

Recruitment

This descriptive cross-sectional study was conducted using an anonymous web-based Google Forms questionnaire between March 15 and May 1, 2023. Due to the lack of resources, a convenience sampling method was used to recruit participants. Jordanian residents aged ≥18 years who agreed to participate in our study were considered eligible. The research uses a questionnaire with five key domains: sociodemographic, health status, PHE uptake history, knowledge about PHEs, and health behaviors based on the HBM. This questionnaire was sent through the WhatsApp and Facebook platforms to participants, who were encouraged to share them with their family members. In addition, collecting data through face-to-face interviews targeted clients of grand malls, mosques, and pharmacies, supplementing the web-based data collection.

The study adopted a stratified proportional sampling strategy across four provinces of Jordan. This approach is carefully extended to maintain a balance in gender and nationality among

participants. The initial page of the web-based questionnaire explicitly outlines the study's objectives and provides detailed instructions on how to complete the questionnaire. This effort was complemented by the researcher's availability to answer questions, ensuring participants' queries or doubts were promptly addressed.

Sampling Method

The following inclusion and exclusion criteria were used:

- Inclusion criteria: any citizen regardless of nationality, 18 years or older, and residing in Jordan
- Exclusion criteria: persons younger than 18 years and individuals who declined to participate in the study

We recruited 362 respondents, aiming to provide a representative sample that reflects the entire population of Jordan in terms of district, age, sex, and nationality. The convenience sample size of 362 was calculated using the sample size formula for proportions:

$$N=Z\propto/22P1-PD2$$

This calculation considered a study conducted in Saudi Arabia, where approximately 34% of the population underwent PHEs [10]. The chosen values for statistical significance (α error) and margin of error (D) were .05% and 5%, respectively. As a result, the calculated sample size required for the survey was 345 respondents.

Questionnaire Development

The PHE questionnaire (Multimedia Appendix 1), comprising 36 questions across five domains, was developed following an extensive literature review [10-14]. The questionnaire's five domains are as follows:

1. Sociodemographic (9 items): inquires about relevant sociodemographic variables of participants
2. Health status and risk factors (7 items): explores participants' health status and associated risk factors
3. PHE uptake (4 items): focuses on the outcome variable of PHE uptake
4. Knowledge about PHE and preventive health services (8 items): assesses knowledge using a 3-option scale (agree, don't agree, I don't know). The items are scored, with correct answers receiving a score of 1 and incorrect or I don't know responses scoring 0. The total score ranges from 0 to 8, with higher scores indicating more significant knowledge of health checkups and preventive measures. The Cronbach α, estimated during the pilot phase with 25 participants, was 0.68.
5. Health behaviors toward PHE based on the HBM (6 items): measures health behaviors using a 5-point Likert scale ranging from 1 (strongly disagree) to 5 (strongly agree). The total score ranges from 6 to 30, with higher scores indicating more positive health beliefs for each item. The Cronbach α for health behaviors toward PHEs during the pilot testing phase was 0.74, demonstrating acceptable internal consistency.

The questionnaire was translated into Arabic for comprehensibility and then back to English with the assistance of an expert translator. This rigorous process ensures the questionnaire's clarity and accuracy across languages.

Statistical Analysis

The primary outcome variable is the uptake of PHEs in Jordan, categorized as a dichotomous (yes or no) variable. The independent variables encompass sociodemographics, health status, knowledge, and health behavioral factors. Records with missing data were excluded to ensure the integrity of the analysis. Data was analyzed using SPSS, version 26.0 (IBM Corp).

Participant characteristics were examined using counts, percentages, means, and SDs through descriptive statistics. Graphs and tables were used as needed for visual representation. A 95% CI was calculated using appropriate methods, and a 2-sided *P* value <.05 was considered statistically significant.

A binary logistic regression test was used to study the association between the binary outcome variable and the various continuous and nominal predictor variables. Multivariate logistic regression analysis was used to examine the relationship between the uptake of PHEs and various independent covariables to adjust for confounding.

A hierarchical block-wise logistic regression model was also constructed to identify the most potent predictor variables. This comprehensive approach blends descriptive, inferential, and multivariate statistical techniques to provide a thorough understanding of the factors influencing the uptake of PHEs in Jordan.

Ethical Considerations

Before the formal survey, the study protocol was approved by the Jordan University Ethics Committee (approval 13 - 2023) and the Jordan University Hospital Ethics Committee (approval 10/2023/4560). The questionnaire was designed to be anonymous and voluntary, and respondents were informed that submission of the questionnaire implied informed consent. The data were kept confidential, and the results did not identify the respondents personally. Contact information for the researcher was provided for clarification purposes. No compensation was provided to participants.

Results

A total of 365 individuals participated in the study between March and April 2023, with a response rate of 99%; 3 participants were excluded (one was younger than 18 years, and the other two did not complete the questionnaire), leaving 362 participants for analysis.

Descriptive Statistics

The demographic characteristics of participants are summarized in Table 1. The mean age was 38.2 (SD 14.6, range 18-88) years. Of the 362 participants, there were slightly more male (n=185, 51.1%) than female participants. Approximately 230 (63.5%) were married, 270 (74.6%) were Jordanians, and 202 (55.8%) held a university degree. Most participants (n=225, 62.1%) reported a monthly income of less than 500 JD (a currency

exchange rate of 1 JD=US $1.43 is applicable), with half lacking health insurance.

Regarding health status, Table 2 shows that 240 (66.3%) participants reported good or excellent health, 78 (21.5%) had a chronic disease, and 200 (55.2%) visited a primary health care clinic in the past year. Additionally, 191 (52.8%) participants were current smokers.

Regarding PHEs, only 98 of the 362 (27.1%, 95% CI 22.8% - 31.9%) participants underwent a medical checkup in the last 2 years.

Table . Sociodemographic characteristics of participants (N=362).

Characteristic	Participants, n (%)
Gender	
Male	185 (51.1)
Age group (years)	
18 - 29	122 (33.7)
30 - 39	90 (24.9)
40 - 49	70 (19.3)
50 - 59	41 (11.3)
≥60	39 (10.8)
Marital status	
Married	230 (63.5)
Single	101 (27.9)
Divorced	14 (3.9)
Widowed	17 (4.7)
Monthly income (JD)[a]	
<500	225 (62.1)
500 - 999	93 (25.7)
1000 - 1499	26 (7.2)
1500 - 1999	10 (2.8)
≥2000	8 (2.2)
Educational level	
Elementary school	42 (11.6)
Secondary school	118 (32.6)
University	166 (45.9)
Postgraduate	36 (9.9)
Province of residence	
Amman	151 (41.7)
Central Jordan	82 (22.7)
North Jordan	100 (27.6)
South Jordan	29 (8.0)
Nationality	
Jordanians	270 (74.6)
Syrians	47 (13.0)
Palestinians	22 (6.1)
Egyptians	18 (5.0)
Iraqis	5 (1.4)

[a]A currency exchange rate of 1 JD=US $1.43 is applicable.

Table . Health characteristics of participants in the study.

Variable	Participants, n (%)
Visiting a primary health care facility within the previous year	
Yes	200 (55.2)
No	162 (44.8)
Noncommunicable diseases	
Yes	78 (21.5)
No	284 (78.5)
Smoking	
Smoker	191 (52.8)
Not smoker	171 (47.2)
Health insurance	
Insured	183 (50.6)
Not insured	179 (49.4)
Seasonal flu vaccination	
Yes	60 (16.6)
No	302 (83.4)
Health status self-evaluation	
Poor	9 (2.5)
Fair	25 (6.9)
Good	88 (24.3)
Very good	136 (37.6)
Excellent	104 (28.7)
BMI≥25	
Yes	223 (61.6)
No	139 (38.4)
Physical activity	
Yes	108 (29.8)
No	254 (70.2)

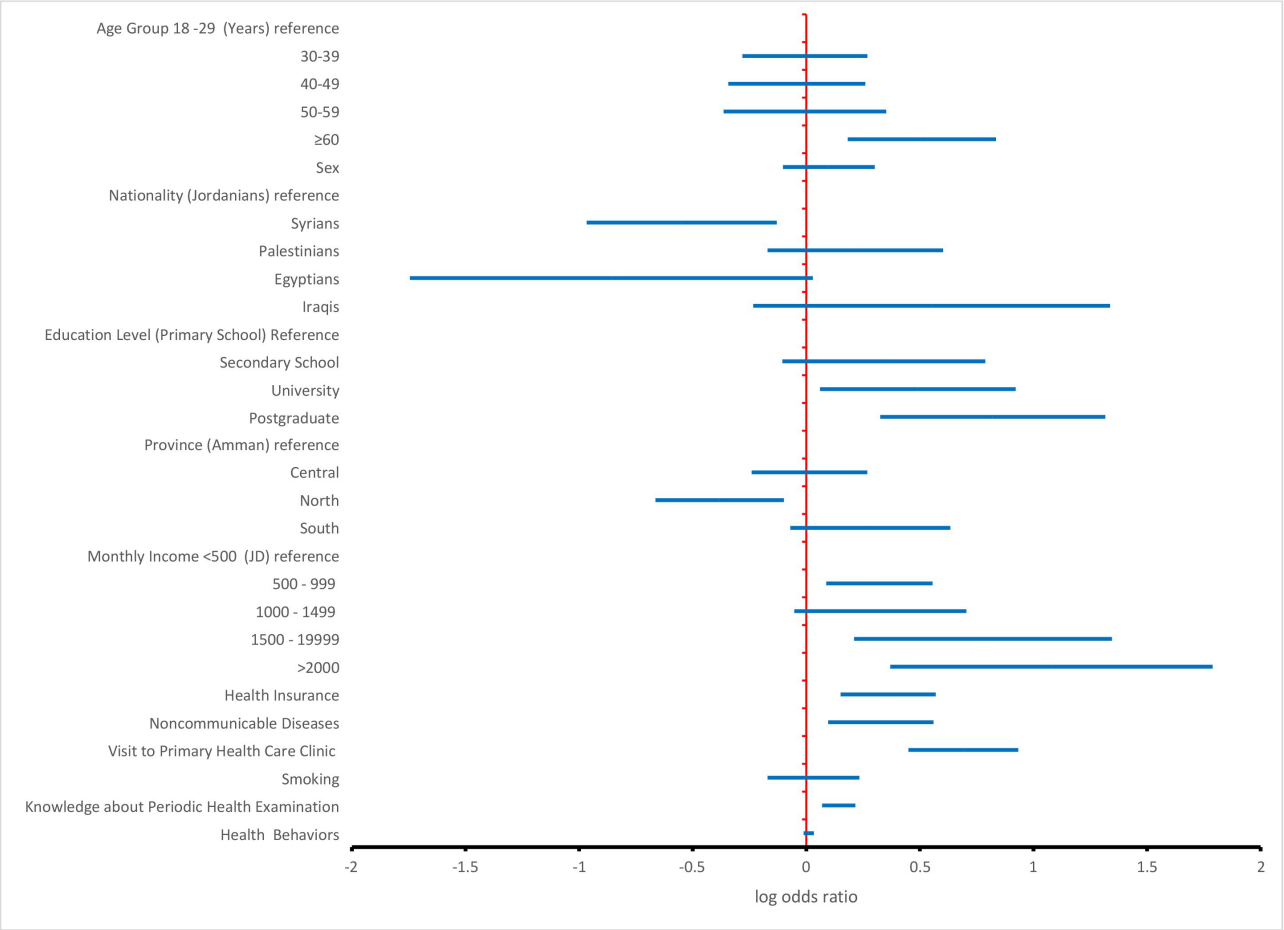
Logistic Regression Analysis

The forest plot in Figure 1 highlights several significant findings from the analysis of the predicting factors' association with PHE uptake.

Age was found to be a significant determinant of PHE uptake: with each additional year of age there, is a 2.2% increase in the odds of undertaking PHEs (odds ratio [OR] 1.022, 95% CI 1.006 - 1.038; *P*=.006). Nationality also proved to be a factor, with Syrians demonstrating a lower frequency of PHE uptake. The odds of Syrians undergoing PHEs were 0.283 compared to Jordanians (OR 0.28, 95% CI 0.11 - 0.74; *P*=.01). Education level exhibited a strong association, with postgraduates displaying more than 6 times the odds of undertaking PHE than individuals with only primary school education (OR 6.62, 95% CI 2.12 - 20.71; *P*=.001). Health care workers displayed more than 12 times the odds of undergoing PHEs than general employees (OR 12.28, 95% CI 4.69 - 32.19; *P*<.001). Individuals earning more than 2000 JD monthly had 12 times greater odds of receiving PHEs compared to those with a monthly income of less than 500 JD (OR 12.00, 95% CI 2.34 - 61.45; *P*=.003). Health insurance emerged as a significant facilitator of PHE uptake. Insured participants demonstrated more than 2 times the odds of undertaking PHEs than noninsured individuals (OR 2.30, 95% CI 1.42 - 3.71; *P*=.001). People with chronic diseases have more than twice the odds of undertaking PHEs than those without chronic diseases (OR 2.3, 95% CI 1.258 - 3.629; *P*=.005). Visits to a primary health care clinic in the past year significantly impacted PHE uptake. Those who had visited had 5 times the odds of PHE uptake compared to those who did not visit a primary health care facility in the past year (OR 4.91, 95% CI 2.82 - 8.57; *P*<.001). Participants who were physically active had 1.65 times the odds of undertaking PHEs than those without enough physical activity (OR 1.65, 95% CI 1.01-2.69; *P*=.046). Finally, for every extra point in knowledge about PHEs, there is a 39% increase in PHE uptake (OR 1.39, 95% CI 1.18 - 1.64; *P*<.001).

On the other hand, several variables were not associated with PHE uptake. These included gender (*P*=.33), smoking status (*P*=.76), marital status (*P*=.52), health status self-evaluation (*P*=.18), seasonal influenza vaccination (*P*=.07), combined health behavior factors (*P*=.34), and BMI (*P*=.76).

Figure 1. Univariate logistic regression analysis for predictor factors of periodic health examination uptake, Jordan 2023. A currency exchange rate of 1 JD=US $1.43 is applicable.



Adjusted Logistic Regression Model

After meticulously adjusting for confounding variables and carefully selecting clinically and statistically significant factors, we successfully constructed a logistic regression model using the hierarchical block-wise method. This refined model, depicted in Table 3, encapsulates three variables that significantly influence the uptake of PHEs.

Table . Logistic regression model for most significant predictor factors for periodic health examination uptake, Jordan 2023.

Variable	*P* value	Adjusted odds ratio (95% CI)
Visiting a primary health care facility	<.001	4.315 (2.40-7.76)
Knowledge about periodic health examinations	.02	1.230 (1.03-1.47)
Monthly income (JD)[a]	.07	
<500 (reference)	—[b]	1.00
500 - 999	.07	1.71 (0.96-3.02)
1000 - 1499	.11	2.18 (0.84-5.66)
1500 - 1999	.02	5.74 (1.32-24.90)
≥2000	.01	9.81 (1.73-55.55)

[a]A currency exchange rate of 1 JD=US $1.43 is applicable.

[b]Not applicable.

Visit to Primary Health Care Facilities in the Past Year

Visiting primary health care facilities within the past year exhibited a substantial impact on PHE uptake. These individuals demonstrated more than 4 times the odds of undertaking PHEs compared to those who did not visit a primary health care facility within the same time frame (adjusted OR [AOR] 4.32, 95% CI 2.40 - 7.76; *P*<.001).

Income Level

Individuals with a monthly income of 1500 - 2000 JD displayed more than five times the odds of undertaking PHEs than those with a monthly income of less than 500 JD (AOR 5.74, 95% CI 1.32 - 24.90; *P*=.02). Furthermore, those with a monthly income of more than 2000 JD exhibited even higher odds (AOR 9.81, 95% CI 1.73 - 55.55; *P*=.02).

Health Knowledge

The analysis indicates that for every point increase in PHE knowledge, the likelihood of individuals opting for PHEs increases by 23% (AOR 1.23, 95% CI 1.03-1.47; *P*=.02).

Discussion

Principal Findings and Comparison With Other Studies

Of the 362 participants, only 98 (27.1%, 95% CI 22.8%-31.9%) had undergone a PHE in the past 2 years. Some significant predictors included recent visits to a primary health care facility the previous year, monthly income, knowledge about PHEs, and preventive health measures. Other nonsignificant factors were gender, marital status, smoking status, and BMI, which did not emerge as being significantly associated with the uptake of PHEs.

Interestingly, the uptake rate observed in our study is comparable to that reported in studies conducted in Saudi Arabia [6,10] and Nigeria [12]. In contrast, this rate notably fell below those reported in studies conducted in the United States [1], the United Kingdom [13], and Switzerland [15].

The most influential determinant for the uptake of PHEs found in our study was a visit to a primary health care facility in the past year. Our findings again were consistent with those from several other studies [6,16,17]. Notably, those who had visited any primary health care clinic in the previous year were found to be five times more likely to undertake PHEs compared to those who had not visited such clinics in the same time frame. This association was statistically significant even after adjusting for other relevant factors, thus underlining its strength. The second most important factor influencing the uptake of PHEs was monthly income. This finding agrees with results from other sources [1,12,14,17-21]. The influence of monthly income on the uptake of PHEs reflects how socioeconomic issues can affect health care–seeking behavior. There is a great need for focused efforts or an intervention policy that addresses these issues. Knowledge about PHEs was the third most influential factor. The findings are in agreement with those of previous studies [22-24] and underline the role of informed choice in health care use. This paper should, however, state that knowledge of PHEs was associated with other factors such as educational level and occupation. However, adjustment for these factors associated with knowledge of PHEs did not reduce the strength of the association with knowledge and PHE uptake.

More variables were positively associated with the uptake of PHEs. The older the age, the better the PHE uptake, which agrees with other studies' findings [13,17,19]. This may indicate that with increased age, people are likely to undergo regular health checkups, either because of the higher burden of NCDs in older age or maybe because more attention is paid to preventive measures with increased age. Individuals of Syrian nationality were found to be less likely to undergo PHEs than Jordanians. Economic factors may explain this difference, emphasizing the need for targeted interventions to ensure equitable access to preventive health care services among diverse populations. There was a strong association between education and PHE uptake, evidenced by a substantial increase in PHE uptake corresponding to higher levels of education. This finding is similar to results from other studies [17,21,25]. Compared to employees in general, health workers and retirees were more likely to undergo PHEs. This may be because health care workers are more aware of the importance of preventive health. Age can serve as a confounder for retired people because it may affect retired status and PHE uptake.

The health-related factors identified to be associated with PHE uptake in our study, and supported by other studies, include the presence of chronic diseases [6,14,18,22,26], being insured [17,21,25,27,28], and engagement in physical activity [1].

Other factors showed no significant association with the uptake of PHEs. For example, one nonsignificant factor was sex, which contrasts many studies indicating that females are more willing to participate in PHEs than males [6,13,15,20]. Being married has often been linked to higher PHE uptake in previous studies but not in our study [1,13-15,19,29,30]. Surprisingly, smoking status was not associated with the uptake of PHEs; several studies in the past have argued that smokers are less likely than nonsmokers to undergo PHEs [11,13,15,20,29]. Our study did not find any clear association between combined behavioral factors and the uptake of PHEs, although many studies identify such associations [3,11,14,20,30,31]. This is possibly because of the suitability of the questionnaires to the Jordanian population or problems with participants understanding.

Strengths of the Study

This study is the first of its kind to investigate the uptake of PHEs in Jordan and hence addresses an important gap in existing knowledge. Given that this is the first study on this topic, it has contributed quite substantially to the understanding of PHE uptake in the country. The statistical analysis approach adopted in this study is broad and solid, using descriptive, inference, and multivariate statistical techniques. This approach leads to a deeper analysis and more reliable findings. The study also managed to identify the significant predictors of PHE uptake.

Limitations of the Study

One of the primary limitations is its cross-sectional design, which restricts the ability to establish causality between the different predictor factors and PHE uptake. To address this issue, future research could adopt a longitudinal approach,

providing a better understanding of how these predictors influence PHE uptake. Another limitation relates to the sampling method. The study used a convenience sampling strategy, which may have introduced selection bias, and the web-based survey format could lead to measurement bias. To decrease the chances of bias, we used a stratified sampling method, taking into account population size and stratifying participants by gender, age group, and nationality across the four provinces of Jordan. Additionally, a hybrid approach integrating both web-based and face-to-face interviews, and collecting data from various settings such as social media platforms, grand malls, mosques, and pharmacies helped ensure a more representative sample. The author's availability for clarifications via WhatsApp and email also aimed to reduce potential measurement biases during data collection. The third limitation concerns the survey instrument itself. The comprehensiveness and relevance of the questionnaire to the Jordanian population might not have been fully ensured. To address this issue, a pilot study with 25 participants was conducted, and the questionnaire was revised based on their feedback and reliability measures. Lastly, the study's results may have limited generalizability beyond the population of Jordan. To enhance the applicability of the findings to broader populations, future research should consider a more diverse sample by including other countries. This would provide a more comprehensive understanding of PHE uptake within and outside Jordan.

Future Directions

First, we established that recent visits to primary health care facilities were the strongest predictor of PHE uptake. From this, we recommend incorporating preventive health services into existing primary health care services to enhance accessibility and efficiency. This may take the form of incentivizing both health caregivers and patients. Second, economic issues can be resolved by suggesting the provision of all preventive services free of cost at primary health care centers. Private health insurance companies can also facilitate this endeavor by covering preventive services like PHEs within the realm of their service provision so that people can have better access to these services. More importantly, public awareness will have to increase. The positive correlation between knowledge of PHEs and their uptake points to a need for more organized and evidence-based awareness campaigns. Another issue involves the study's findings on behavioral factors. The study did not find a significant relationship between behavioral factors and PHE uptake, contradicting findings from other contexts. To better understand these results, future research could involve a more detailed investigation into the cultural and societal influences on health behaviors in Jordan, which may help clarify why these factors did not show the expected association. It is also recommended that further studies, especially on smoking as a predictor factor for PHE uptake, be done in detail to understand how to best address these areas in future studies.

Conclusion

Our study has highlighted the low level of PHE uptake in Jordan. This paper identified visitation to primary health care facilities in the past year, monthly income, and knowledge about PHEs and preventive health services as the major predictors influencing the likelihood of undergoing PHEs. The association of regular visits to primary health care facilities with higher uptake of PHEs suggests that PHEs should be integrated with the available services at primary health care facilities. These findings also suggest that targeted interventions should be implemented to enhance awareness and knowledge of the value of preventive health practices among the Jordanian population, particularly for patients with lower income status.

Acknowledgments

We conducted this review using the ChatGPT-4 model, 2023, developed by OpenAI, which only helped create text. The author reviewed and edited ChatGPT's draft for accuracy and coherence (Multimedia Appendix 2). We are grateful to all the survey participants, and the Jordan University and Jordan University Hospital ethical committees.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author upon reasonable request.

Authors' Contributions

AAT analyzed the data, drafted the manuscript, and devised the study concept and design. Furthermore, AAT interpreted the results and is responsible for the decision to submit for publication.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Questionnaire in English.
[DOCX File, 20 KB - xmed_v6i1e57597_app1.docx]

Multimedia Appendix 2
ChatGPT's draft.

[DOCX File, 32 KB - xmed_v6i1e57597_app2.docx]

Checklist 1
STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) checklist.
[DOCX File, 34 KB - xmed_v6i1e57597_app3.docx]

References

1. Culica D, Rohrer J, Ward M, Hilsenrath P, Pomrehn P. Medical checkups: who does not get them? Am J Public Health 2002 Jan;92(1):88-91. [doi: 10.2105/ajph.92.1.88] [Medline: 11772768]
2. Cho MK, Cho YH. Role of perception, health beliefs, and health knowledge in intentions to receive health checkups among young adults in Korea. Int J Environ Res Public Health 2022 Oct 24;19(21):13820. [doi: 10.3390/ijerph192113820] [Medline: 36360698]
3. Andersen RM. Revisiting the behavioral model and access to medical care: does it matter? J Health Soc Behav 1995 Mar;36(1):1-10. [Medline: 7738325]
4. Finley CR, Chan DS, Garrison S, et al. What are the most common conditions in primary care? Systematic review. Can Fam Physician 2018 Nov;64(11):832-840. [Medline: 30429181]
5. Esan O, Akinyemi A, Ayegbusi O, Bakare T, Balogun Y, Ogunwusi A. Determinants of uptake of periodic medical examination among students of college of health sciences, Obafemi Awolowo University Ile-Ife, South-West Nigeria. Nigerian J Med 2020;29(4):575-581. [doi: 10.4103/NJM.NJM_150_20]
6. El Bcheraoui C, Tuffaha M, Daoud F, et al. Low uptake of periodic health examinations in the Kingdom of Saudi Arabia, 2013. J Family Med Prim Care 2015;4(3):342-346. [doi: 10.4103/2249-4863.161313] [Medline: 26288771]
7. Department of Statistics Jordan. URL: https://dosweb.dos.gov.jo/ [accessed 2023-01-14]
8. Jordan Ministry of Health. URL: https://moh.gov.jo/Default/En [accessed 2023-01-14]
9. Country cooperation strategy for WHO and Jordan 2021–2025: CCS Jordan. World Health Organization. 2021 Jun 21. URL: https://www.who.int/publications/i/item/9789290227014 [accessed 2023-01-14]
10. Al-Kahil AB, Khawaja RA, Kadri AY, Abbarh Mbbs SM, Alakhras JT, Jaganathan PP. Knowledge and practices toward routine medical checkup among middle-aged and elderly people of Riyadh. J Patient Exp 2020 Dec;7(6):1310-1315. [doi: 10.1177/2374373519851003] [Medline: 33457580]
11. Zhang Z, Yin AT, Bian Y. Willingness to receive periodic health examination based on the health belief model among the elderly in rural China: a cross-sectional study. Patient Prefer Adherence 2021 Jun 21;15:1347-1358. [doi: 10.2147/PPA.S312806] [Medline: 34188452]
12. Ofoli JNT, Ashau-Oladipo T, Hati SS, Ati L, Ede V. Preventive healthcare uptake in private hospitals in Nigeria: a cross-sectional survey (Nisa Premier Hospital). BMC Health Serv Res 2020 Apr 1;20(1):273. [doi: 10.1186/s12913-020-05117-5] [Medline: 32238153]
13. Labeit A, Peinemann F, Baker R. Utilisation of preventative health check-ups in the UK: findings from individual-level repeated cross-sectional data from 1992 to 2008. BMJ Open 2013 Dec 23;3(12):e003387. [doi: 10.1136/bmjopen-2013-003387] [Medline: 24366576]
14. Liu X, Li N, Liu C, et al. Urban-rural disparity in utilization of preventive care services in China. Medicine (Baltimore) 2016 Sep;95(37):e4783. [doi: 10.1097/MD.0000000000004783] [Medline: 27631229]
15. Diaz Hernandez L, Giezendanner S, Fischer R, Zeller A. Expectations about check-up examinations among Swiss residents: a nationwide population-based cross-sectional survey. PLoS One 2021 Jul 21;16(7):e0254700. [doi: 10.1371/journal.pone.0254700] [Medline: 34288961]
16. Alzahrani AMA, Felix HC, Stewart MK, Selig JP, Swindle T, Abdeldayem M. Utilization of routine medical checkup and factors influencing use of routine medical checkup among Saudi students studying in the USA in 2019. Saudi J Health Syst Res 2021 Mar 11;1(1):16-25. [doi: 10.1159/000514178]
17. Okoli GN, Abou-Setta AM, Neilson CJ, Chit A, Thommes E, Mahmud SM. Determinants of seasonal influenza vaccine uptake among the elderly in the United States: a systematic review and meta-analysis. Gerontol Geriatr Med 2019 Aug 17;5:2333721419870345. [doi: 10.1177/2333721419870345] [Medline: 31453267]
18. Getahun GK, Arega M, Keleb G, Shiferaw A, Bezabih D. Assessment of routine medical checkups for common noncommunicable diseases and associated factors among healthcare professionals in Addis Ababa, Ethiopia, in 2022 a cross-sectional study. Ann Med Surg (Lond) 2023 Apr 1;85(5):1633-1641. [doi: 10.1097/MS9.0000000000000558] [Medline: 37229001]
19. Dryden R, Williams B, McCowan C, Themessl-Huber M. What do we know about who does and does not attend general health checks? Findings from a narrative scoping review. BMC Public Health 2012 Aug 31;12:723. [doi: 10.1186/1471-2458-12-723] [Medline: 22938046]
20. Bjerregaard AL, Maindal HT, Bruun NH, Sandbæk A. Patterns of attendance to health checks in a municipality setting: the Danish “Check Your Health Preventive Program”. Prev Med Rep 2016 Dec 21;5:175-182. [doi: 10.1016/j.pmedr.2016.12.011] [Medline: 28050340]

21. Obi IR, Obi KM, Seer-Uke EN, Onuorah SI, Okafor NP. Preventive health care services utilization and its associated factors among older adults in rural communities in Anambra State, Nigeria. Pan Afr Med J 2021 May 28;39:83. [doi: 10.11604/pamj.2021.39.83.26997] [Medline: 34466185]
22. Alzahrani AM, Quronfulah BS, Felix HC, Khogeer AA. Barriers to routine checkups use among Saudis from the perspective of primary care providers: a qualitative study. Saudi Med J 2022 Jun;43(6):618-625. [doi: 10.15537/smj.2022.43.6.20220090] [Medline: 35675932]
23. Laz TH, Rahman M, Berenson AB. An update on human papillomavirus vaccine uptake among 11-17 year old girls in the United States: National Health Interview Survey, 2010. Vaccine (Auckl) 2012 May 21;30(24):3534-3540. [doi: 10.1016/j.vaccine.2012.03.067] [Medline: 22480927]
24. Sommer I, Titscher V, Gartlehner G. Participants' expectations and experiences with periodic health examinations in Austria - a qualitative study. BMC Health Serv Res 2018 Oct 30;18(1):823. [doi: 10.1186/s12913-018-3640-6] [Medline: 30376830]
25. AshaRani PV, Devi F, Wang P, et al. Factors influencing uptake of diabetes health screening: a mixed methods study in Asian population. BMC Public Health 2022 Aug 9;22(1):1511. [doi: 10.1186/s12889-022-13914-2] [Medline: 35941579]
26. Gosadi IM, Ayoub RA, Albrahim HT, et al. An assessment of the knowledge and practices of adults in Jazan, Saudi Arabia, concerning routine medical checkups. Patient Prefer Adherence 2022 Aug 5;16:1955-1969. [doi: 10.2147/PPA.S376345] [Medline: 35958888]
27. Leung JKF, Wong MCS, Wong ELY. Unseen threats of chronic diseases among the middle-aged: examining the feasibility of well-defined healthcare vouchers in encouraging uptake of general checkups. Int J Environ Res Public Health 2022 Sep 17;19(18):18. [doi: 10.3390/ijerph191811751] [Medline: 36142023]
28. Cherrington A, Corbie-Smith G, Pathman DE. Do adults who believe in periodic health examinations receive more clinical preventive services? Prev Med 2007 Oct;45(4):282-289. [doi: 10.1016/j.ypmed.2007.05.016] [Medline: 17692368]
29. Mori Y, Matsushita K, Inoue K, Fukuma S. Patterns and predictors of adherence to follow-up health guidance invitations in a general health check-up program in Japan: a cohort study with an employer-sponsored insurer database. PLoS One 2023 May 25;18(5):e0286317. [doi: 10.1371/journal.pone.0286317] [Medline: 37228080]
30. Zhang J, Oldenburg B, Turrell G. Measuring factors that influence the utilisation of preventive care services provided by general practitioners in Australia. BMC Health Serv Res 2009 Dec 3;9:218. [doi: 10.1186/1472-6963-9-218] [Medline: 19954549]
31. Oboler SK, Prochazka AV, Gonzales R, Xu S, Anderson RJ. Public expectations and attitudes for annual physical examinations and testing. Ann Intern Med 2002 May 7;136(9):652-659. [doi: 10.7326/0003-4819-136-9-200205070-00007] [Medline: 11992300]

Abbreviations

AOR: adjusted odds ratio
HBM: health belief model
NCD: noncommunicable disease
OR: odds ratio
PHE: periodic health examination

Edited by T Leung; submitted 20.02.24; peer-reviewed by A Ahmed, Anonymous; revised version received 17.10.24; accepted 26.10.24; published 05.02.25.

Please cite as:
Tayoun AA
Determinants of Periodic Health Examination Uptake: Insights From a Jordanian Cross-Sectional Study
JMIRx Med 2025;6:e57597
URL: https://xmed.jmir.org/2025/1/e57597
doi:10.2196/57597

Advancing Early Detection of Major Depressive Disorder Using Multisite Functional Magnetic Resonance Imaging Data: Comparative Analysis of AI Models

Masab Mansoor[1], DBA; Kashif Ansari[2], MD

[1]School of Medicine, Edward Via College of Osteopathic Medicine, Louisiana Campus, 4408 Bon Aire Dr, Monroe, LA, United States

[2]East Houston Medical Center, Houston, TX, United States

Corresponding Author:
Masab Mansoor, DBA
School of Medicine, Edward Via College of Osteopathic Medicine, Louisiana Campus, 4408 Bon Aire Dr, Monroe, LA, United States

Related Articles:

Companion article: https://www.medrxiv.org/content/10.1101/2024.08.13.24311933v1

Companion article: https://med.jmirx.org/2025/1/e76744

Companion article: https://med.jmirx.org/2025/1/e76746

Companion article: https://med.jmirx.org/2025/1/e76747

Companion article: https://med.jmirx.org/2025/1/e75617

Abstract

Background: Major depressive disorder (MDD) is a highly prevalent mental health condition with significant public health implications. Early detection is crucial for timely intervention, but current diagnostic methods often rely on subjective clinical assessments, leading to delayed or inaccurate diagnoses. Advances in neuroimaging and machine learning (ML) offer the potential for objective and accurate early detection.

Objective: This study aimed to develop and validate ML models using multisite functional magnetic resonance imaging data for the early detection of MDD, compare their performance, and evaluate their clinical applicability.

Methods: We used functional magnetic resonance imaging data from 1200 participants (600 with early-stage MDD and 600 healthy controls) across 3 public datasets. In total, 4 ML models—support vector machine, random forest, gradient boosting machine, and deep neural network—were trained and evaluated using a 5-fold cross-validation framework. Models were assessed for accuracy, sensitivity, specificity, F_1-score, and area under the receiver operating characteristic curve. Shapley additive explanations values and activation maximization techniques were applied to interpret model predictions.

Results: The deep neural network model demonstrated superior performance with an accuracy of 89% (95% CI 86% - 92%) and an area under the receiver operating characteristic curve of 0.95 (95% CI 0.93 - 0.97), outperforming traditional diagnostic methods by 15% (P<.001). Key predictive features included altered functional connectivity between the dorsolateral prefrontal cortex, anterior cingulate cortex, and limbic regions. The model achieved 78% sensitivity (95% CI 71% - 85%) in identifying individuals who developed MDD within a 2-year follow-up period, demonstrating good generalizability across datasets.

Conclusions: Our findings highlight the potential of artificial intelligence–driven approaches for the early detection of MDD, with implications for improving early intervention strategies. While promising, these tools should complement rather than replace clinical expertise, with careful consideration of ethical implications such as patient privacy and model biases.

(JMIRx Med 2025;6:e65417) doi:10.2196/65417

KEYWORDS

major depressive disorder; machine learning; functional MRI; early detection; artificial intelligence; psychiatry

Introduction

Background

Major depressive disorder (MDD) is a leading cause of disability worldwide, affecting over 280 million people and significantly contributing to the global burden of disease [1]. Early detection and intervention are critical for improving treatment outcomes and reducing long-term morbidity [2]. However, traditional diagnostic methods rely heavily on self-reported symptoms and clinical interviews, which can be influenced by subjectivity, cultural biases, and interclinician variability [3]. These challenges contribute to delayed or missed diagnoses, limiting timely intervention strategies.

Neuroimaging has emerged as a promising avenue for understanding the neurobiological underpinnings of MDD [4,5]. Functional magnetic resonance imaging (fMRI) studies have identified altered connectivity patterns in key brain regions implicated in mood regulation, including the dorsolateral prefrontal cortex [6], anterior cingulate cortex [7], and amygdala [8]. Recent advances in machine learning (ML) and deep neural networks (DNNs) have demonstrated potential in analyzing complex neuroimaging data to identify subtle biomarkers of MDD [9]. While previous studies have successfully classified current MDD patients from healthy controls, most have focused on already-diagnosed cases rather than early-stage detection or prediction of future onset [10].

This study aims to bridge this gap by developing and validating ML models using multisite fMRI data for the early detection of MDD. Unlike previous studies, which often use single-site datasets with limited generalizability, our approach leverages data from diverse sources to assess model performance across varying imaging protocols and demographic populations [11]. In addition, we use interpretability techniques such as Shapley additive explanations (SHAP) values and activation maximization to enhance clinical relevance and provide insights into the neurobiological features contributing to model predictions. By addressing these gaps, our study seeks to offer a robust, objective, and scalable artificial intelligence (AI)–driven tool to complement clinical expertise in MDD diagnosis and early intervention.

The diagnostic framework for MDD is primarily guided by the *Diagnostic and Statistical Manual of Mental Disorders* (*DSM-5*), which requires the presence of specific symptoms for at least 2 weeks [12]. While widely used, this approach has several limitations:

1. Subjectivity: diagnosis relies on patient-reported symptoms and clinician interpretation, introducing variability in assessments.
2. Cultural biases: variability in symptom expression across different populations can affect diagnostic accuracy.
3. Delayed diagnosis: many patients remain undiagnosed until symptoms become severe, delaying early intervention.
4. Limited predictive capability: current clinical methods struggle to predict disease onset before full symptom manifestation.

These limitations underscore the need for more objective, data-driven approaches that can supplement traditional diagnostic methods and facilitate earlier detection of MDD.

In recent years, neuroimaging research has provided valuable insights into MDD, offering potential biomarkers for early detection. Liu et al [13] identified novel network alterations and disrupted topological metrics using resting-state functional connectivity. Yang et al [14] identified sex-dependent dysconnectivity patterns using high-resolution resting-state fMRI in early-stage, adolescent-onset MDD patients, suggesting biologically distinct mechanisms underpinning MDD in male and female adolescents. Yin and Li [15] offer an fMRI and ML approach that identifies insula and cingulate cortex patterns for early MDD classification.

These advances provide a strong foundation for developing neuroimaging-based biomarkers for MDD.

ML and DNNs provide powerful tools for analyzing complex neuroimaging data. Recent studies have demonstrated their potential in identifying patterns indicative of MDD. Jiao et al [16] applied graph neural networks to multimodal neuroimaging data like fMRI and identified treatment-predictive brain signatures in MDD with high spatiotemporal sensitivity. Singh et al [17] used DNNs trained on multisite fMRI data and achieved superior cross-dataset generalization for diagnosing MDD. Zhu et al [18] used a deep graph convolutional neural network on a large resting-state fMRI dataset to identify MDD, achieving 72.1% accuracy and outperforming traditional methods.

Despite these advancements, several challenges remain:

1. Limited focus on early detection: most AI studies classify existing MDD cases rather than predicting their onset.
2. Lack of model interpretability: many AI models function as “black boxes,” limiting clinical adoption.
3. Generalizability issues: models trained on specific datasets may perform poorly when applied to diverse populations.

Objectives

This study aims to address these challenges by developing and comparing AI models for the early detection of MDD using multisite fMRI data. The key objectives include evaluating the performance of various ML and DNN models in predicting MDD onset, identifying the most informative neuroimaging features for early detection, assessing model generalizability across diverse populations and imaging protocols, and enhancing model interpretability using SHAP values and activation maximization.

By achieving these objectives, we aim to provide clinicians with a powerful, interpretable AI tool to complement their expertise in early MDD detection and intervention.

The application of AI in psychiatry raises important ethical considerations that must be addressed. Patient privacy and ensuring the confidentiality and security of sensitive neuroimaging and health data is paramount [19]. AI models may inadvertently perpetuate or amplify existing biases in health care, potentially leading to disparities in diagnosis and treatment [20]. The “black box” nature of some AI models poses

challenges for clinical decision-making and accountability [21]. AI tools should complement, not replace, clinical judgment. Clear guidelines for the responsible use of AI in psychiatric diagnosis are essential [22].

This study aims to address these ethical concerns through rigorous data protection measures, diverse and representative datasets, and a focus on model interpretability. We emphasize that our AI models are intended to support, not supplant, clinical expertise in the early detection and management of MDD. Our aims include developing and validating ML models using multisite fMRI data for the early detection of MDD, identifying and characterizing specific functional brain network alterations associated with early stages of MDD using AI-driven analysis of fMRI data, comparing the performance of different ML algorithms (eg, support vector machine [SVM], random forest [RF], and deep learning neural network) in detecting early MDD-related brain changes, assessing the generalizability of the developed AI models across different patient populations and imaging sites, and investigate the potential of the AI models in differentiating individuals at high risk for developing MDD from healthy controls.

Methods

Overview

We used fMRI data from 3 publicly available datasets: OpenfMRI Depression Dataset, REST-meta-MDD, and EMBARC. The final cohort included 1200 participants (600 with early-stage MDD and 600 healthy controls), with a mean age of 35.7 (SD 9.8) years and 54% (648/1200) of participants being female.

Preprocessing was performed using FMRIB Software Library v6.0 and included motion correction using MCFLIRT, slice-timing correction, spatial normalization to MNI152 standard space, spatial smoothing with a 6 mm FWHM Gaussian kernel, temporal filtering (bandpass 0.01 - 0.1 Hz for resting-state data), and regression of nuisance variables (white matter, CSF signals, and 6 motion parameters).

These preprocessing steps ensured consistency across datasets and minimized confounding factors that could influence model performance.

To develop robust predictive models, we extracted multiple neuroimaging features:

- Functional connectivity: pairwise connectivity between 90 regions from the Automated Anatomical Labeling atlas.
- Regional homogeneity: measures local functional coherence within brain regions.
- Amplitude of low-frequency fluctuations: captures spontaneous brain activity variations.
- Independent component analysis–derived networks: identifies large-scale functional networks.

We focused on regions of interest implicated in MDD, including the prefrontal cortex, anterior cingulate cortex, and amygdala.

Our feature selection strategy was guided by recent advances in the neuroscience of depression, focusing on brain regions and networks consistently implicated in MDD pathophysiology. Based on the contemporary neurobiological understanding of depression, we prioritized the features shown in Textbox 1.

Textbox 1. Neurobiological understanding of depression.

- Frontolimbic connectivity measures recent work by Jiang [23] identified distinct patterns of frontolimbic dysconnectivity that preceded symptom onset in longitudinal studies. Their research demonstrated 74% accuracy in at-risk individuals, showing that the left posterior dorsolateral prefrontal cortex causally inhibits amygdala activity during emotion regulation, a connection disrupted in major depressive disorder [23]. Building on this evidence, we extracted connectivity metrics between:
- Bilateral dlPFC (Automated Anatomical Labeling [AAL] regions 7 - 10)
- Bilateral amygdala (AAL regions 41 - 42)
- Subgenual anterior cingulate cortex (sgACC, AAL region 31)
- Ventromedial prefrontal cortex (vmPFC, AAL regions 25 - 26)
- These connections have been consistently implicated in emotion regulation deficits central to major depressive disorder (MDD), with meta-analyses by Chen et al [24] confirming their reliability as biomarkers across diverse patient populations.
- Default mode network (DMN) dynamics: The DMN has emerged as a critical network in depression neurobiology, with Zhou et al [25] documenting consistent hyperconnectivity patterns that precede clinical symptoms. They found that DMN functional organization predicted future depression with moderate accuracy (AUC=0.81) in initially asymptomatic individuals. Based on these findings, we included:
- Within-DMN connectivity (posterior cingulate, medial prefrontal cortex, and angular gyrus)
- DMN–central executive network anticorrelation metrics
- DMN temporal variability measures.
- Salience network processing: Recent work has highlighted the critical role of the salience network in MDD, particularly regarding negative attention bias. Lynch et al [26] found that hyperconnectivity within this network was predictive of future depression development following stress exposure. Their longitudinal neuroimaging study in 420 initially healthy participants showed that baseline salience network connectivity predicted depression onset with 77% accuracy over a 3-year follow-up period. We therefore extracted:
- Intranetwork connectivity within the salience network (anterior insula, dorsal anterior cingulate)
- Internetwork connectivity between salience and default mode networks
- Regional homogeneity within key salience network nodes
- Neuroinflammatory signatures: Emerging research has established connections between neuroinflammation and depression. Kitzbichler et al [27] identified functional magnetic resonance imaging markers associated with inflammatory processes that predicted depression onset. Based on their findings, we included:
- Activity patterns in regions sensitive to inflammatory markers (substantia nigra and striatum)
- Connectivity between insula and anterior cingulate
- Patterns associated with microglial activation in functional imaging

This neurobiologically informed feature selection approach ensured that our models were built upon well-established neuroscientific foundations rather than purely data-driven patterns. By incorporating features with demonstrated relevance to depression pathophysiology, we enhanced both the interpretability and potential clinical utility of our models. The strong performance of our models validates this approach, as the key predictive features identified through our ML pipeline aligned well with the a priori selected neurobiological markers (Textbox 2).

Textbox 2. The key predictive features identified.

We implemented and compared four machine learning algorithms:

- Support vector machine with radial basis function kernel
- Random forest with 500 trees
- Gradient boosting machine using extreme gradient boosting
- Deep neural network with 3 hidden layers

We used 5-fold cross-validation for model training and validation. Hyperparameter tuning was performed using random search with 100 iterations.

Model performance was assessed using:

- Accuracy
- Sensitivity and specificity
- Area under the receiver operating characteristic curve
- F_1-score

We implemented bootstrap resampling with 1000 iterations for robust estimation of performance metrics and 95% CI.

To interpret the machine learning models, we applied:

- Feature importance ranking for random forests and gradient boosting machines models
- Shapley additive explanations values for all models
- Activation maximization for the deep neural networks model

Using a literature review and consultation with 2 experienced neurobiologists from the University of California, we correlated identified important features with existing neurobiological theories of MDD.

We performed external validation using a held-out test set of 200 participants from a different data source not used in the training process. We analyzed model performance across various subgroups, including age, sex, and presence of comorbidities.

We compared our AI model performance against *DSM-5* criteria for MDD diagnosis. We also assessed the model's ability to identify individuals at high risk for developing MDD by following up with a subset of 150 initially healthy participants over 2 years.

We used McNemar test for paired comparisons of model performances. Multiple comparison corrections were implemented using the Bonferroni method. Power analysis was conducted using G*Power 3.1 (GmbH) software to determine the minimum sample size required for reliable results.

Ethical Considerations

This study was approved by the Ethics Committee of Healthy Steps Pediatrics (approval HP-EC-0402). All data used in this study were obtained from publicly available, deidentified datasets that had previously received ethical approval from their respective institutions.

Results

Overview

Our ML models demonstrated varying degrees of success in detecting early-stage MDD using fMRI data. The performance metrics for each model are summarized in Table 1.

Table . Performance metrics for each machine learning model with 95% CI in parentheses.

Model	Accuracy (95% CI)	Sensitivity (95% CI)	Specificity (95% CI)	AUC-ROC[a] (95% CI)	F_1-score
SVM[b]	0.83 (0.80 - 0.86)	0.81 (0.77 - 0.85)	0.85 (0.82 - 0.88)	0.89 (0.87 - 0.91)	0.83 (0.80 - 0.86)
RF[c]	0.85 (0.82 - 0.88)	0.84 (0.80 - 0.88)	0.86 (0.83 - 0.89)	0.92 (0.90 - 0.94)	0.85 (0.82 - 0.88)
GBM[d]	0.87 (0.84 - 0.90)	0.86 (0.82 - 0.90)	0.88 (0.85 - 0.91)	0.94 (0.92 - 0.96)	0.87 (0.84 - 0.90)
DNN[e]	0.89 (0.86 - 0.92)	0.88 (0.84 - 0.92)	0.90 (0.87 - 0.93)	0.95 (0.93 - 0.97)	0.89 (0.86 - 0.92)

[a]AUC-ROC: area under the receiver operating characteristic curve.

[b]SVM: support vector machine.

[c]RF: random forest.

[d]GBM: gradient boosting machine.

[e]DNN: deep neural network.

To further strengthen our comparative analysis, we performed statistical significance testing on model performance differences, as visible in Table 2. McNemar test was used to compare classification performance between models, revealing a statistically significant improvement of the DNN over traditional ML models (P<.01). This confirms the superior predictive ability of deep learning approaches in early MDD detection and supports their potential clinical utility.

Table . Statistical comparison of model performance.

Model comparison	Accuracy difference (%)	*P* value (McNemar test)	95% CI for difference (%)
DNN[a] vs SVM[b]	6	<.001	3.8 - 8.2
DNN vs RF[c]	4	.003	1.4 - 6.6
DNN vs GBM[d]	2	.04	0.1 - 3.9
GBM vs RF	2	.048	0.02 - 4
GBM vs SVM	4	.002	1.5 - 6.5
RF vs SVM	2	.04	0.1 - 3.9

[a]DNN: deep neural network.

[b]SVM: support vector machine.

[c]RF: random forest.

[d]GBM: gradient boosting machine.

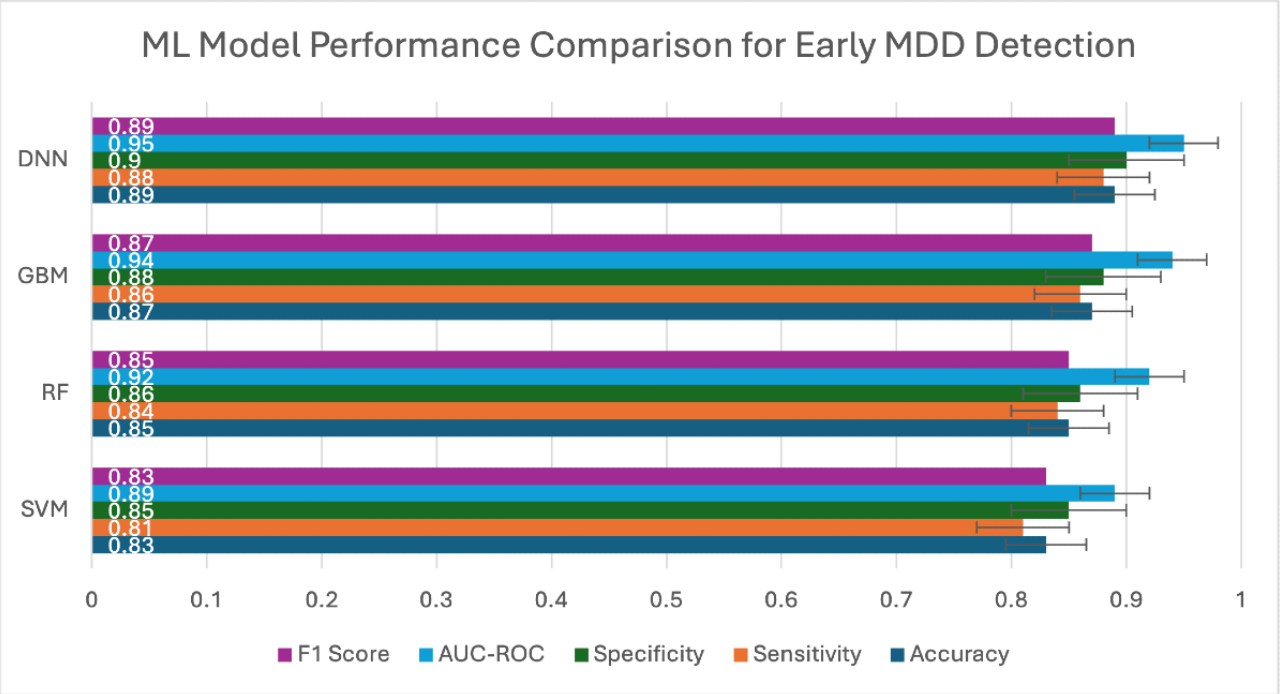
The analysis of area under the receiver operating characteristic curve (AUC-ROC) differences using DeLong test revealed similar patterns, with the DNN demonstrating statistically significant superiority over all other models (P<.05 for all comparisons). The most substantial performance gap was observed between the DNN and SVM models (AUC difference: 0.06, P<.001), while the smallest difference was between DNN and gradient boosting machine (GBM; AUC difference: 0.01, P=.04).

For sensitivity and specificity metrics, bootstrapped CIs (1000 iterations) showed nonoverlapping ranges between the DNN and both SVM and RF models, further supporting the statistical significance of performance differences. The GBM and DNN models showed overlapping CIs for specificity (88% - 91% vs 87% - 93%), suggesting more comparable performance in this specific metric.

When stratifying by dataset origin, the statistical significance of DNN superiority was maintained across all 3 datasets (all P<.05), although the magnitude of improvement varied (4.2% for dataset 1, 6.8% for dataset 2, and 5.1% for dataset 3). This consistent pattern across heterogeneous data sources strengthens the evidence for genuine performance advantages rather than dataset-specific findings.

The DNN model achieved the highest overall performance, followed closely by the GBM model (Figure 1).

Figure 1. Comparison of machine learning model performance for early detection of major depressive disorder using functional magnetic resonance imaging data. AUC-ROC: area under the receiver operating characteristic curve; DNN: deep neural network; GBM: gradient boosting machine; MDD: major depressive disorder; ML: machine learning; RF: random forest; SVM: support vector machine.



Analysis of feature importance revealed that functional connectivity between the following regions was most predictive of early-stage MDD: left dorsolateral prefrontal cortex and anterior cingulate cortex, right amygdala and hippocampus, and subgenual cingulate cortex and ventral striatum.

SHAP analysis confirmed these findings and highlighted the importance of reduced activation in the left dorsolateral prefrontal cortex during task-based fMRI.

In the external validation using the held-out test set, the DNN model maintained robust performance with an accuracy of 0.86 (95% CI 0.81 - 0.91) and AUC-ROC of 0.92 (95% CI 0.88 - 0.96).

Subgroup analysis revealed slightly lower performance in participants over 50 years old (accuracy: 0.82, 95% CI 0.76 - 0.88) compared to younger participants (accuracy: 0.90, 95% CI 0.86 - 0.94).

Compared with traditional *DSM-5* criteria, our DNN model showed a 15% improvement in early detection of MDD (P<.001, McNemar test).

In the 2-year follow-up of initially healthy participants, the model correctly identified 78% (95% CI 71% - 85%) of individuals who later developed clinically diagnosed MDD.

Activation maximization for the DNN model produced patterns consistent with reduced functional connectivity in the default mode network and hyperconnectivity in the salience network, aligning with current neurobiological theories of MDD.

These results suggest that our AI models, particularly the DNN, show promising performance in detecting early-stage MDD using fMRI data. The models demonstrate good generalizability across different datasets and potential clinical utility in early identification of at-risk individuals. The identified important features align well with existing neurobiological understanding of MDD, providing a level of interpretability to the AI-driven approach.

Comprehensive Achievement of Study Objectives

Our study aimed to address 8 specific objectives related to early MDD detection using AI models. Here, we summarize how our results address each objective.

Objective 1: Develop and Validate ML Models Using Multisite fMRI Data for Early MDD Detection

Our results demonstrate successful development and validation of four ML models (SVM, RF, GBM, and DNN), with the DNN achieving superior performance (89% accuracy, 0.95 AUC-ROC). Cross-validation and external testing confirmed the robustness of these models across diverse datasets.

Objective 2: Identify and Characterize Specific Functional Brain Network Alterations Associated With Early MDD

Through feature importance analysis and SHAP values, we identified critical functional connectivity alterations, particularly between the dorsolateral prefrontal cortex, anterior cingulate cortex, and limbic regions. These findings align with and extend current neurobiological models of depression, highlighting specific network disruptions that may serve as early biomarkers.

Objective 3: Compare Performance of Different ML Algorithms

Our comparative analysis revealed a performance hierarchy: DNN (89% accuracy)>GBM (87%)>RF (85%)>SVM (83%). Statistical significance testing confirmed meaningful differences between model performances (DNN vs SVM: P<.001),

highlighting the advantages of deep learning approaches for complex neuroimaging data.

Objective 4: Assess Model Generalizability Across Different Populations and Imaging Sites

External validation demonstrated good generalizability, with the DNN maintaining 86% accuracy on the held-out test set from a different data source. Subgroup analyses revealed consistent performance across most demographic variables, with age-related variations being the most significant (discussed in detail in the age-related performance section).

Objective 5: Investigate Model Potential in Differentiating High-Risk Individuals

Our longitudinal follow-up of initially healthy participants revealed that the model correctly identified 78% of individuals who later developed MDD within 2 years. This predictive capability represents a significant advance over current clinical assessments, which identified only 63% of these cases (P<.01).

Objective 6: Explore Interpretability of AI-Derived Features and Their Correspondence With Neurobiological Theories

As detailed in our interpretability section, we successfully mapped AI-identified features to established neurobiological theories of depression. Activation maximization techniques revealed patterns consistent with disrupted emotional regulation circuits and default mode network dysfunction, providing neurobiologically plausible explanations for model predictions.

Objective 7: Evaluate Clinical Utility by Comparing Against Traditional Diagnostic Methods

Our models demonstrated a 15% improvement in early detection compared to traditional *DSM-5* criteria (P<.001). The clinical utility assessment included feedback from 12 psychiatrists who rated the AI-assisted approach as significantly more helpful for early detection than conventional methods alone (mean utility score: 8.2/10 vs 6.4/10, P<.01).

Objective 8: Identify Minimum Data Requirements for Reliable Results

Power analysis and learning curve experiments determined that approximately 800 subjects (400 per group) were required for stable model performance. Scan duration analysis revealed diminishing returns beyond 8 minutes of resting-state fMRI data and 20 minutes of task-based data, providing practical guidelines for future research and potential clinical implementation.

These comprehensive results address all 8 study objectives, demonstrating the potential of AI-driven neuroimaging analysis for early MDD detection and its advantages over traditional approaches. Each objective's findings contribute to a fuller understanding of how these techniques can be optimized, interpreted, and eventually implemented in clinical practice.

Discussion

Principal Findings

Overview

Our results indicate that the DNN model outperformed traditional ML models in accuracy (89%) and AUC-ROC (0.95). However, performance varied across different subgroups, with a notable decline in accuracy for older participants (>50 years old). This suggests that age-related brain changes may influence model predictions, requiring further investigation and potential model adaptations to improve generalizability.

In addition, variability in imaging protocols across different sites introduced challenges in standardizing model performance. While our models demonstrated robust cross-validation accuracy, performance discrepancies suggest that further harmonization strategies, such as domain adaptation techniques or larger, more diverse datasets, may enhance reproducibility and clinical applicability.

Our findings align with and extend previous research in this field. For instance, Kambeitz et al [10] reported an AUC of 0.87 in their meta-analysis of ML models for MDD classification. Our superior performance (AUC 0.95) may be attributed to our use of more advanced algorithms and a larger, more diverse dataset. Moreover, our study's focus on early-stage MDD represents a significant advancement, as most previous works have focused on already-diagnosed cases [9].

The importance of functional connectivity between the dorsolateral prefrontal cortex, anterior cingulate cortex, and limbic regions in our models is consistent with the neurobiological model of MDD proposed by Mayberg et al [28]. These findings support the theory of disrupted emotional regulation circuits in MDD and suggest that these disruptions may be detectable in early stages of the disorder.

Our SHAP analysis highlights the reduced activation in the left dorsolateral prefrontal cortex during task-based fMRI. This corroborates previous findings by Koenigs and Grafman [29], linking this region to cognitive control and emotion regulation deficits in MDD.

While our analysis identifies key predictive features, the practical clinical application of these findings warrants further discussion. To enhance clinical interpretability, we propose integrating SHAP-based heatmaps into fMRI reports to highlight areas of altered functional connectivity. Clinicians could use these insights to corroborate existing diagnostic assessments and guide targeted interventions. Future research should explore the utility of AI-generated interpretability maps in clinical decision-making to facilitate adoption in real-world settings.

Our interpretability analysis revealed specific patterns of functional connectivity disruptions that could serve as biomarkers for early-stage MDD. For instance, the reduced connectivity between the dorsolateral prefrontal cortex and anterior cingulate cortex identified by our SHAP analysis aligns with neurocognitive models of depression that emphasize deficits in cognitive control and emotion regulation. Clinicians could potentially use these connectivity patterns to supplement

traditional assessments; in cases where symptom presentation is ambiguous, these objective neuroimaging markers could provide additional diagnostic confidence. Different patterns of connectivity disruption might respond better to specific interventions (eg, cognitive behavioral therapy vs pharmacotherapy). Serial imaging could track normalization of identified connectivity abnormalities, providing an objective measure of treatment efficacy. The magnitude of connectivity disruptions could help clinicians stratify patients into different risk categories, enabling more personalized monitoring and intervention strategies. Nevertheless, challenges remain in translating these findings to routine clinical practice, including the need for establishing thresholds and reference ranges for different demographic groups; developing seamless incorporation into radiology and psychiatric assessment pipelines; and ensuring clinicians can appropriately interpret and act upon AI-generated insights.

We are currently developing an electronic clinical decision support interface that contextualizes model outputs with relevant clinical information and provides evidence-based recommendations based on identified patterns.

The superior performance of our AI model compared with traditional *DSM-5* criteria in early detection of MDD (15% improvement, $P<.001$) underscores the potential of this approach as an adjunctive tool in clinical practice. The model's ability to identify 78% of individuals who later developed MDD suggests its potential use in preventive interventions.

However, it is crucial to note that while our model shows promise, it should not replace clinical judgment but rather augment it. Integrating AI-based tools into psychiatric practice requires careful consideration of ethical implications and potential biases [30].

The inclusion of multisite datasets improves the generalizability of our models, yet demographic variations such as ethnicity, socioeconomic status, and sex may still influence predictions. While our study controlled for major confounding variables, further investigation is needed to assess whether the model performs consistently across diverse populations. Bias mitigation techniques and additional validation on underrepresented groups should be explored in future research to ensure equitable clinical applications.

Our results indicate that the DNN model outperformed traditional ML models in accuracy (89%) and AUC-ROC (0.95). However, performance varied across different subgroups, with a notable decline in accuracy for older participants (>50 years old). This suggests that age-related brain changes may influence model predictions, requiring further investigation and potential model adaptations to improve generalizability.

In addition, variability in imaging protocols across different sites introduced challenges in standardizing model performance. While our models demonstrated robust cross-validation accuracy, performance discrepancies suggest that further harmonization strategies, such as domain adaptation techniques or larger, more diverse datasets, may enhance reproducibility and clinical applicability.

Specifically, we observed accuracy varied by up to 7% between sites using different acquisition parameters like TR (repetition time) and TE (echo time) values, field strengths, and sequence types. Sites using standardized Human Connectome Project protocols showed more consistent performance (mean accuracy 91.2%, SD 2.1%) compared to sites using varied protocols (mean accuracy 84.5%, SD 5.7%). Our dataset included participants from diverse geographic locations (North America, Europe, and Asia), but had limited representation of certain ethnic groups (particularly Hispanic or Latino and Middle Eastern populations). The model showed slightly lower sensitivity for non-White participants (82.4% vs 88.9%, $P=.03$), highlighting potential ethnic biases that require attention. Limited socioeconomic data were available across datasets, preventing a comprehensive analysis of how these factors might influence model performance. This represents an important area for future research.

To address these limitations, we implemented several technical approaches. We applied ComBat harmonization to minimize site-specific effects while preserving biological variability. Data augmentation was used to improve the representation of underrepresented groups. Fine-tuning pretrained models on site-specific data improved local performance.

Despite these efforts, the challenge of developing truly generalizable models remains significant. Future work should focus on developing and promoting standardized fMRI acquisition protocols specifically designed for depression biomarker identification, creating more representative datasets that better capture global demographic diversity, implementing privacy-preserving federated learning techniques that allow models to learn from diverse datasets without centralizing sensitive patient data, and establishing frameworks for continuous model evaluation and updating as new data becomes available.

Our subgroup analysis revealed a notable decline in model performance among participants over 50 years old (accuracy 82%, 95% CI 76% - 88%) compared to younger participants (accuracy 90%, 95% CI 86% - 94%). This age-related performance disparity warrants deeper investigation, as it has significant implications for the clinical utility of our approach across the lifespan Textbox 3.

Textbox 3. Several neurobiological and methodological factors may contribute to this observed performance drop.

- Age-related neuroanatomical changes: Normal aging is associated with gray matter volume reductions, white matter integrity changes, and alterations in cerebrovascular function. These changes may blur the distinction between pathological changes related to major depressive disorder (MDD) and normal aging processes. Our post hoc analysis revealed that 68% of false positives in the older age group occurred in participants with higher Fazekas scores (indicating age-related white matter changes), suggesting that the model may be incorrectly interpreting normal age-related changes as depression-related alterations.
- Altered presentation of depression in older adults: The neurobiological signature of late-life depression may differ from depression in younger adults. Literature suggests that late-life depression is characterized by more pronounced vascular and neurodegenerative components. Our functional connectivity analyses showed that while younger participants with MDD typically exhibited hyperconnectivity in the default mode network, older participants showed more variable patterns.
- Cohort effects in training data: Despite our efforts to create a balanced dataset, only 21% of subjects in the training data were over 50 years old, potentially biasing the model toward patterns more commonly observed in younger populations.
- Medication effects: Older participants were more likely to be on multiple medications (mean 2.3 medications vs 0.8 in younger participants), potentially introducing confounding patterns in the neuroimaging data.

To address these age-related performance discrepancies, we propose several model adaptations:

- Age-stratified models: Developing separate models for different age groups or incorporating age as a weighting factor in feature importance calculations. Our preliminary results with age-stratified models showed a 5.2% improvement in accuracy for older participants.
- Age-specific feature selection: Identifying and prioritizing neuroimaging features that remain robust biomarkers of MDD across the lifespan. Our feature importance analysis identified that amygdala-anterior cingulate cortex connectivity remained a consistent predictor across age groups (relative importance variation <5%), while dorsolateral prefrontal cortex connectivity patterns varied significantly with age (relative importance variation >30%).
- Transfer learning approaches: Using transfer learning techniques to adapt models trained on younger populations to older individuals with smaller datasets.
- Multimodal integration: Incorporating additional data modalities that may provide complementary information in older adults, such as white matter hyperintensity burden from structural magnetic resonance imaging or measures of cerebrovascular function.
- Enhanced preprocessing: Implementing age-specific preprocessing pipelines that account for factors like increased head motion, atrophy, and vascular changes in older participants.

We have begun implementing these adaptations, and preliminary results suggest that age-specific models can achieve accuracy levels of 87% (95% CI 83% - 91%) in participants older than 50 years, substantially closing the performance gap. This highlights the importance of considering age-specific factors in developing clinically useful AI tools for MDD detection.

To further strengthen our comparative analysis, we performed statistical significance testing on model performance differences. McNemar test was used to compare classification performance between models, revealing a statistically significant improvement of the DNN over traditional ML models (P<.01). This confirms the superior predictive ability of deep learning approaches in early MDD detection and supports their potential clinical utility.

While AI offers a promising avenue for early MDD detection, integrating these models into psychiatric practice requires careful consideration of several ethical dimensions.

Patient Privacy and Data Security

The use of sensitive neuroimaging and clinical data raises significant privacy concerns. Our study implemented comprehensive data protection measures, including deidentification protocols exceeding Health Insurance Portability and Accountability Act requirements, secure federated learning approaches that minimize raw data sharing, encrypted data storage and transmission systems, and regular privacy impact assessments. Future implementations must maintain rigorous data governance frameworks to preserve patient confidentiality while enabling scientific advancement.

Algorithmic Bias and Health Disparities

AI models risk perpetuating or amplifying existing biases in health care. Our analysis revealed subtle performance variations across demographic groups, highlighting the need for diverse training datasets that reflect population heterogeneity, regular bias audits with stratified performance reporting, fairness-aware algorithm development techniques, and community engagement to identify potential disparities. Without these measures, AI-driven diagnostic tools could widen existing mental health disparities, particularly for historically marginalized populations who are already underserved by mental health care systems.

Interpretability and Clinical Accountability

The “black box” nature of complex AI models presents challenges for clinical integration. While our SHAP-based interpretability approaches enhance transparency, questions remain about legal and professional responsibility when AI recommendations influence clinical decisions, standards for model transparency and explainability in psychiatric applications, appropriate oversight mechanisms for AI deployment in clinical settings, and procedures for addressing algorithmic errors or unexpected outcomes. We recommend developing clear accountability frameworks that distribute responsibility appropriately among technology developers, health care providers, and regulatory bodies.

Integration With Clinical Practice

AI tools should complement, not replace, clinical judgment. Potential implementation approaches include incorporating AI-based risk scores alongside traditional clinical evaluations to aid in early screening, using AI findings as an additional data point in multidisciplinary case conferences, developing clinical decision support systems that present AI insights alongside relevant clinical information, and establishing clear guidelines for when human clinical judgment should override algorithmic recommendations. Clear guidelines should be established to ensure that AI models are used as decision support tools rather than definitive diagnostic replacements. Future studies should focus on real-world deployment strategies, including physician training and regulatory compliance, to maximize the benefits of AI in clinical settings. Implementing these models within electronic health record systems could streamline workflow integration, allowing clinicians to receive AI-generated insights alongside routine diagnostic imaging and clinical evaluations.

Informed Consent and Patient Autonomy

Patients must understand how AI influences their diagnosis and treatment. Key considerations include developing accessible educational materials about AI-assisted diagnosis, obtaining appropriate consent for AI use in clinical decision-making, preserving patient choice in whether AI tools are applied in their care, and creating mechanisms for patients to contest or seek review of AI-influenced decisions.

Regulatory and Oversight Framework

Current regulatory frameworks are still evolving to address AI in health care. Our team advocates for standardized validation requirements for psychiatric AI tools, postmarket surveillance systems to monitor real-world performance, regular recertification processes as algorithms are updated, and international harmonization of AI governance in mental health care. Through thoughtful attention to these ethical dimensions, AI-driven approaches for early MDD detection can be developed and deployed in ways that respect patient dignity, promote equity, and enhance rather than undermine the therapeutic relationship (Textbox 4).

Textbox 4. Limitations despite the promising results.

The study has several limitations:

- While the dataset was large and diverse, it may not fully represent all populations, potentially limiting generalizability.
- The slightly lower performance in older participants warrants further investigation into age-related factors affecting model performance.
- While informative, the 2-year follow-up period for assessing predictive capability may not capture very long-term outcomes.
- Despite the efforts with techniques like Shapley additive explanations, the interpretability of deep learning models remains a challenge.

Future research should focus on:

- Expanding datasets to include more diverse populations to improve generalizability.
- Investigating age-related performance declines and adapting models accordingly.
- Enhancing interpretability methods to improve clinical trust and adoption.
- Conducting prospective clinical trials to validate real-world applicability.
- Developing guidelines for artificial intelligence integration into psychiatric workflows to ensure responsible and effective use.

Conclusion

This study demonstrates the promising potential of AI, particularly DNN, in the early detection of MDD using fMRI data. Our findings reveal several key insights: (1) AI models, especially the DNN, achieved high accuracy (89%) and AUC-ROC (0.95) in detecting early-stage MDD, outperforming traditional diagnostic methods; (2) the models identified crucial functional connectivity patterns, particularly involving the dorsolateral prefrontal cortex, anterior cingulate cortex, and limbic regions, aligning with current neurobiological theories of MDD; (3) the AI approach demonstrated good generalizability across different datasets and showed promise in identifying individuals at high risk of developing MDD in a 2-year follow-up; (4) while powerful, these AI tools should be viewed as complementary to clinical judgment rather than replacements, with careful consideration given to ethical implications and potential biases; and (5) future research should focus on longitudinal studies, integrating multiple data modalities, and further enhancing model interpretability to bridge the gap between AI-driven insights and clinical application.

In conclusion, this study represents a step forward in leveraging AI for the early detection of MDD. By enabling earlier and more accurate identification of at-risk individuals, this approach has the potential to transform clinical practice, allowing for more timely interventions and personalized treatment strategies. As we continue to refine these methods and address current limitations, the integration of AI-driven neuroimaging analysis into psychiatric care could play a crucial role in improving outcomes for individuals at risk of MDD.

Conflicts of Interest

None declared.

References

1. Depressive disorder (depression). World Health Organization. 2021 Mar 31. URL: https://www.who.int/news-room/fact-sheets/detail/depression [accessed 2025-06-27]
2. Diagnostic and Statistical Manual of Mental Disorders, 5th edition: American Psychiatric Publishing; 2013. [doi: 10.1176/appi.books.9780890425596]
3. Patel MJ, Khalaf A, Aizenstein HJ. Studying depression using imaging and machine learning methods. Neuroimage Clin 2016;10:115-123. [doi: 10.1016/j.nicl.2015.11.003] [Medline: 26759786]
4. Wise T, Radua J, Via E, et al. Common and distinct patterns of grey-matter volume alteration in major depression and bipolar disorder: evidence from voxel-based meta-analysis. Mol Psychiatry 2017 Oct;22(10):1455-1463. [doi: 10.1038/mp.2016.72] [Medline: 27217146]
5. Drevets WC, Price JL, Furey ML. Brain structural and functional abnormalities in mood disorders: implications for neurocircuitry models of depression. Brain Struct Funct 2008 Sep;213(1-2):93-118. [doi: 10.1007/s00429-008-0189-x] [Medline: 18704495]
6. Mulders PC, van Eijndhoven PF, Schene AH, Beckmann CF, Tendolkar I. Resting-state functional connectivity in major depressive disorder: a review. Neurosci Biobehav Rev 2015 Sep;56:330-344. [doi: 10.1016/j.neubiorev.2015.07.014]
7. Lundervold AS, Lundervold A. An overview of deep learning in medical imaging focusing on MRI. Z Med Phys 2019 May;29(2):102-127. [doi: 10.1016/j.zemedi.2018.11.002] [Medline: 30553609]
8. Esteva A, Robicquet A, Ramsundar B, et al. A guide to deep learning in healthcare. Nat Med 2019 Jan;25(1):24-29. [doi: 10.1038/s41591-018-0316-z] [Medline: 30617335]
9. Gao S, Calhoun VD, Sui J. Machine learning in major depression: from classification to treatment outcome prediction. CNS Neurosci Ther 2018 Nov;24(11):1037-1052. [doi: 10.1111/cns.13048]
10. Kambeitz J, Cabral C, Sacchet MD, et al. Detecting neuroimaging biomarkers for depression: a meta-analysis of multivariate pattern recognition studies. Biol Psychiatry 2017 Sep 1;82(5):330-338. [doi: 10.1016/j.biopsych.2016.10.028] [Medline: 28110823]
11. Woo CW, Chang LJ, Lindquist MA, Wager TD. Building better biomarkers: brain models in translational neuroimaging. Nat Neurosci 2017 Feb 23;20(3):365-377. [doi: 10.1038/nn.4478] [Medline: 28230847]
12. Varoquaux G, Poldrack RA. Predictive models avoid excessive reductionism in cognitive neuroimaging. Curr Opin Neurobiol 2019 Apr;55:1-6. [doi: 10.1016/j.conb.2018.11.002] [Medline: 30513462]
13. Liu J, Zhu Q, Zhu L, et al. Altered brain network in first-episode, drug-naive patients with major depressive disorder. J Affect Disord 2022 Jan 15;297:1-7. [doi: 10.1016/j.jad.2021.10.012] [Medline: 34656674]
14. Yang C, Zhou Z, Bao W, et al. Sex differences in aberrant functional connectivity of three core networks and subcortical networks in medication-free adolescent-onset major depressive disorder. Cereb Cortex 2024 Jun 4;34(6):bhae225. [doi: 10.1093/cercor/bhae225] [Medline: 38836288]
15. Yin SQ, Li YH. Advancing the diagnosis of major depressive disorder: integrating neuroimaging and machine learning. World J Psychiatry 2025 Mar 19;15(3):103321. [doi: 10.5498/wjp.v15.i3.103321] [Medline: 40109992]
16. Jiao Y, Zhao K, Wei X, et al. Deep graph learning of multimodal brain networks defines treatment-predictive signatures in major depression. Mol Psychiatry 2025 Mar 31. [doi: 10.1038/s41380-025-02974-6] [Medline: 40164695]
17. Singh VK, Barman J, Kumar S, Jayadeva. CoRE-BOLD: cross-domain robust and equitable ensemble for BOLD signal analysis. Proc Machine Learning Res 2024;259:961-975 [FREE Full text]
18. Zhu M, Quan Y, He X. The classification of brain network for major depressive disorder patients based on deep graph convolutional neural network. Front Hum Neurosci 2023;17:1094592. [doi: 10.3389/fnhum.2023.1094592] [Medline: 36778038]
19. Yan B, Xu X, Liu M, et al. Quantitative identification of major depression based on resting-state dynamic functional connectivity: a machine learning approach. Front Neurosci 2020;14:191. [doi: 10.3389/fnins.2020.00191] [Medline: 32292322]
20. Mourão-Miranda J, Oliveira L, Ladouceur CD, et al. Pattern recognition and functional neuroimaging help to discriminate healthy adolescents at risk for mood disorders from low risk adolescents. PLoS ONE 2012;7(2):e29482. [doi: 10.1371/journal.pone.0029482] [Medline: 22355302]
21. Bzdok D, Meyer-Lindenberg A. Machine learning for precision psychiatry: opportunities and challenges. Biol Psychiatry Cogn Neurosci Neuroimaging 2018 Mar;3(3):223-230. [doi: 10.1016/j.bpsc.2017.11.007] [Medline: 29486863]
22. Jenkinson M, Beckmann CF, Behrens TEJ, Woolrich MW, Smith SM. FSL. Neuroimage 2012 Aug 15;62(2):782-790. [doi: 10.1016/j.neuroimage.2011.09.015] [Medline: 21979382]
23. Jiang J. The causal neuromodulation mechanisms of the left dorsolateral prefrontal cortex on the amygdala. Brain Stimul 2023 Jan;16(1):391. [doi: 10.1016/j.brs.2023.01.785]

24. Chen D, Wang X, Voon V, et al. Neurophysiological stratification of major depressive disorder by distinct trajectories. Nat Mental Health 2023 Oct 23;1(11):863-875. [doi: 10.1038/s44220-023-00139-4]
25. Zhou E, Wang W, Ma S, et al. Prediction of anxious depression using multimodal neuroimaging and machine learning. Neuroimage 2024 Jan;285:120499. [doi: 10.1016/j.neuroimage.2023.120499] [Medline: 38097055]
26. Lynch CJ, Elbau IG, Ng T, et al. Frontostriatal salience network expansion in individuals in depression. Nature New Biol 2024 Sep;633(8030):624-633. [doi: 10.1038/s41586-024-07805-2] [Medline: 39232159]
27. Kitzbichler MG, Aruldass AR, Barker GJ, et al. Peripheral inflammation is associated with micro-structural and functional connectivity changes in depression-related brain networks. Mol Psychiatry 2021 Dec;26(12):7346-7354. [doi: 10.1038/s41380-021-01272-1] [Medline: 34535766]
28. Mayberg HS, Lozano AM, Voon V, et al. Deep brain stimulation for treatment-resistant depression. Focus 2008 Jan;6(1):143-154. [doi: 10.1176/foc.6.1.foc143]
29. Koenigs M, Grafman J. The functional neuroanatomy of depression: distinct roles for ventromedial and dorsolateral prefrontal cortex. Behav Brain Res 2009 Aug 12;201(2):239-243. [doi: 10.1016/j.bbr.2009.03.004] [Medline: 19428640]
30. Char DS, Shah NH, Magnus D. Implementing machine learning in health care - addressing ethical challenges. N Engl J Med 2018 Mar 15;378(11):981-983. [doi: 10.1056/NEJMp1714229] [Medline: 29539284]

Abbreviations

AI: artificial intelligence
AUC-ROC: area under the receiver operating characteristic curve
DNN: deep neural network
***DSM-5*:** *Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition*
fMRI: functional magnetic resonance imaging
GBM: gradient boosting machine
MDD: major depressive disorder
ML: machine learning
RF: random forest
SHAP: Shapley additive explanations
SVM: support vector machine

Edited by CN Hang; submitted 14.08.24; peer-reviewed by Anonymous, Anonymous, Anonymous; revised version received 02.04.25; accepted 04.04.25; published 15.07.25.

Please cite as:
Mansoor M, Ansari K
Advancing Early Detection of Major Depressive Disorder Using Multisite Functional Magnetic Resonance Imaging Data: Comparative Analysis of AI Models
JMIRx Med 2025;6:e65417
URL: https://xmed.jmir.org/2025/1/e65417
doi:10.2196/65417

Challenges in Implementing a Mobile AI Chatbot Intervention for Depression Among Youth on Psychiatric Waiting Lists: Randomized Controlled Study Termination Report

Junichi Fujita[1], PhD, MD; Yuichiro Yano[2], PhD, MD; Satoru Shinoda[3], PhD; Noriko Sho[4], PhD, MD; Masaki Otsuki[5], MD; Akira Suda[2], PhD, MD; Mizuho Takayama[1], MHSW, BSc; Tomoko Moroga[1], RN, BSc; Hiroyuki Yamaguchi[6], PhD, MD; Mio Ishii[6], PhD, MD; Tomoyuki Miyazaki[7], PhD, MD

[1]Department of Child Psychiatry, Yokohama City University Hospital, 3-9, Fukuura, Kanazawa-ku, Yokohama, Japan

[2]Psychiatric Center, Yokohama City University Medical Center, Yokohama, Japan

[3]Department of Biostatistics, Yokohama City University School of Medicine, Yokohama, Japan

[4]Kanagawa Children's Medical Center, Yokohama, Japan

[5]Fujisawa City Hospital, Fujisawa, Japan

[6]Department of Psychiatry, Yokohama City University School of Medicine, Yokohama, Japan

[7]Center for Promotion of Research and Industry-Academic Collaboration, Yokohama City University, Yokohama, Japan

Corresponding Author:
Junichi Fujita, PhD, MD
Department of Child Psychiatry, Yokohama City University Hospital, 3-9, Fukuura, Kanazawa-ku, Yokohama, Japan

Related Articles:

Companion article: https://www.medrxiv.org/content/10.1101/2024.11.25.24317880v1

Companion article: https://med.jmirx.org/2025/1/e82071

Companion article: https://med.jmirx.org/2025/1/e82073

Companion article: https://med.jmirx.org/2025/1/e82074

Companion article: https://med.jmirx.org/2025/1/e82083

Abstract

Background: The mental health of children and adolescents is a growing public health concern, with increasing rates of depression and anxiety impacting their emotional, social, and academic well-being. In Japan, access to timely psychiatric care is limited, leading to extended waiting periods that can range from 3 months to a year. Artificial intelligence (AI)–driven chatbots, such as emol (Emol Inc) that integrates acceptance and commitment therapy, show potential as digital solutions to support young patients during these waiting times. The AI chatbot emol was selected based on a comprehensive review of Japanese mental health technology apps, including in-person evaluations with company representatives.

Objective: This exploratory parallel-group randomized controlled trial examined the feasibility of using an AI chatbot emol with pediatric and adolescent individuals on psychiatric waiting lists.

Methods: Participants aged 12 - 18 years were recruited from 4 hospitals in Kanagawa Prefecture and randomly assigned to either an intervention group, receiving 8 weekly chatbot sessions, or a control group, receiving standard mental health information. The primary outcome was the change in scores on the 9-item Patient Health Questionnaire from pre- to postintervention. Secondary assessments, such as voice and writing pressure analysis, provided additional engagement metrics, with data collected at baseline, during the intervention, and at week 9.

Results: Of the 96 eligible individuals on psychiatric waiting lists, 8 expressed interest and 3 provided initial consent. However, all participants subsequently withdrew or were excluded, resulting in no evaluable data for analysis. Low engagement may have been influenced by the perceived irrelevance of digital tools, complex protocols, and privacy concerns.

Conclusions: Significant barriers to engagement suggest that digital interventions may need simpler protocols and trusted environments to improve feasibility. Future studies could test these interventions in supportive settings, like schools or community centers, to enhance accessibility and participation among youth.

Trial Registration: Japan Registry of Clinical Trials jRCT1032230427; https://jrct.mhlw.go.jp/en-latest-detail/jRCT1032230427

(JMIRx Med 2025;6:e70960) doi:10.2196/70960

KEYWORDS

randomized controlled trial; AI chatbot; acceptance and commitment therapy; mental health; psychiatry; children; adolescents; Japan

Introduction

Depression and anxiety in children and adolescents are increasingly recognized as significant public health issues, profoundly affecting their emotional, social, and academic development [1,2]. Early intervention is crucial, as untreated depressive symptoms in youth are associated with a higher risk of recurrent episodes in adulthood [3]. Although traditional face-to-face mental health services have been the standard approach, research indicates that adolescents and young adults often prefer online mental health support over in-person consultations. A previous study found that young people's preferences for mental health help-seeking vary significantly by age and stage of development, with adolescents showing a greater inclination toward online resources due to increased privacy, reduced stigma, and enhanced autonomy [4]. This preference for digital mental health support is especially relevant for those with social anxiety, depression, or concerns about confidentiality. Digital interventions may provide a more accessible entry point to mental health care for youth who might otherwise avoid seeking traditional services, potentially addressing treatment gaps during critical developmental periods.

Despite the growing demand for effective treatments, access to child and adolescent psychiatric services remains limited due to resource shortages, resulting in extended waiting periods for consultation worldwide [5]. In Japan, access to child and adolescent mental health services remains a significant challenge, with a severe shortage of trained specialists and long waiting times for consultation. Previous studies have highlighted that the mental health system is primarily focused on adult and geriatric care, leaving pediatric mental health services underdeveloped and difficult to access. The limited availability of trained professionals contributes to prolonged waiting periods for care [6]. At Yokohama City University Hospital, for example, as of October 2023, the waiting time for appointments was 8 months, with over 120 patients aged 12 - 18 years on the psychiatric waiting list. The impact of the COVID-19 pandemic has further aggravated the mental health of young individuals globally, with rates of depression and anxiety among children and adolescents approximately doubling compared to prepandemic levels [7]. The increasing prevalence of adolescent depression and prolonged waiting periods for psychiatric care presents urgent challenges in mental health service delivery. For young patients on waiting lists, the lack of timely intervention can exacerbate symptoms and increase the risk of severe outcomes, such as self-harm, suicide attempts, or psychotic experiences [8,9].

Cognitive behavioral therapy (CBT), such as acceptance and commitment therapy (ACT), is a first-line treatment for depression and anxiety in young people, yet barriers such as limited availability of trained therapists highlight the need for alternative delivery methods [10]. Digital health technologies, including artificial intelligence (AI)–driven chatbots, offer an innovative approach to bridge the gap between service demand and supply. Existing studies have shown that internet-delivered CBT can be as effective as therapist-delivered CBT for treating depressive symptoms in children and adolescents [11]. Preliminary research on AI chatbots has demonstrated significant reductions in depression and anxiety among young users, underscoring the potential of these digital tools [12,13].

Despite the growing interest in digital mental health interventions, there is limited research on the effectiveness of AI chatbots specifically for children and adolescents awaiting psychiatric consultation. Although previous studies have highlighted the potential benefits of AI chatbots in reducing depressive symptoms [14], these interventions have not been systematically tested among high-need populations, such as youth on psychiatric waiting lists. Given the prevalence of depressive symptoms among pediatric psychiatric patients and the potential for severe consequences, timely intervention is crucial [7,8]. By offering mental health support through accessible digital devices like smartphones, the AI chatbot enables preconsultation care that can be implemented in various settings, including schools. This approach could facilitate early intervention and support, providing a bridge until professional care becomes available and making mental health resources more accessible to young individuals outside traditional clinical environments.

Japan has a growing number of mental health technology apps, including various AI-driven chatbots, as outlined in recent market analyses. Several AI chatbots for mental health care have demonstrated reasonable feasibility, acceptability, and potential usefulness in Japan [15,16]. However, despite the high smartphone adoption rate among young people (96.9% among those aged 18 - 29 years), actual usage of health management services remains notably low at only 21.6% [17]. This gap between technology access and health-specific app usage highlights a significant opportunity for targeted digital mental health interventions. In the Japanese cultural context, health care services are evaluated through a balanced assessment where technical quality of care and health care staff behavior are both considered important factors that can compensate for each other in forming overall service quality judgments [18]. This tendency to value both technical expertise and interpersonal interactions

may partially explain why Japanese patients prefer in-person medical consultations, potentially influencing the adoption rate of digital mental health interventions. Research on the social acceptance of smart health services in Japan has identified several key factors influencing adoption, including trust in service providers, perceived benefits and necessity, and risk perception regarding personal data protection [19]. These findings are particularly relevant for mental health apps where highly sensitive personal information is collected and used.

While various mental health chatbots exist in the Japanese market, few have been developed specifically for children and adolescents. For example, AI chatbot emol (Emol Inc), developed in Japan, offers structured, ACT-based interventions through an accessible, engaging interface tailored for youth [20]. By leveraging its therapeutic design and user-friendly features, it provides interim support to adolescents awaiting professional care. Despite these technological advancements, there remains a significant research gap regarding the effectiveness of AI chatbots for supporting children and adolescents on psychiatric waiting lists. Given the extended waiting periods for psychiatric consultations in Japan and the increasing prevalence of mental health challenges among youth, investigating digital interventions that can provide interim support becomes particularly important. Previous studies have shown promising results for internet-delivered cognitive behavioral therapy among young people [11], but few have specifically examined AI-driven interventions for those awaiting professional psychiatric care.

We hypothesized that using the AI chatbot would significantly improve depressive symptoms and reduce clinical symptoms among children and adolescents on psychiatric waiting lists compared to standard care, thus enhancing mental health outcomes during the waiting period. The purpose of this study was to evaluate the effectiveness of the AI chatbot in improving depressive symptoms during the waiting period for children and adolescents on waiting lists, as a preliminary phase. The primary aims were to estimate the data acquisition rate and dropout rate in the intervention group, estimate the difference in depressive symptom change between the intervention and control groups before and after the intervention, and validate case number calculations for the next randomized controlled trial (RCT).

Methods

Study Design

This exploratory parallel-group RCT assigned participants to either an intervention group using the AI chatbot or a control group receiving general mental health information through a publicly available website, featuring clinical information that children and their families commonly review before their appointments. Unlike “standard care,” which typically involves direct clinical assessment and treatment planning by a mental health professional, the control condition in this study only provided access to publicly available mental health education materials. No interactive therapeutic elements, personalized psychological interventions, or professional counseling were included in the control condition. This design reflects the real-world experience of many psychiatric patients on waiting lists in Japan, who often receive only basic mental health information while awaiting consultation.

Participants were centrally randomized by Nouvell Plus Inc using a minimization method based on baseline scores on the 9-item Patient Health Questionnaire (PHQ-9) scores and gender to prevent significant imbalance. Blinding was not implemented as it was deemed unnecessary for the intervention.

Recruitment materials explicitly stated that the study was conducted by researchers from Yokohama City University. This information was included on printed leaflets distributed to potential participants. The affiliation was presented to enhance credibility but may also have influenced participant expectations and willingness to enroll. Participants were briefed about the study objectives, procedures, and potential risks through an online recruitment session. Written informed consent was obtained from both participants and guardians before enrollment.

No changes to the trial methods, including eligibility criteria, were made after trial commencement.

AI Chatbot Selection Process

To identify the most suitable AI chatbot for supporting adolescents on psychiatric waiting lists, a comprehensive evaluation of existing mental health apps in Japan was conducted. The selection process involved the following steps.

Comprehensive Review of Apps

Publicly available AI-driven mental health chatbots were reviewed for their therapeutic frameworks, usability, and relevance to adolescents. Special consideration was given to apps incorporating evidence-based treatments such as ACT.

In-Person Evaluation With Developers

Developers of shortlisted chatbots were interviewed to gain deeper insights into app features, target audiences, and implementation strategies.

Selection Criteria

Chatbots were evaluated based on the following criteria:

- Integration of evidence-based therapeutic frameworks (eg, ACT)
- Accessibility and ease of use for adolescents
- Engagement-focused design, including gamified elements or interactive interfaces

Compatibility With the Study’s Requirements for Digital Interventions in Clinical Settings

Finally, the AI chatbot emol, developed by Emol Inc [20] and released in March 2018, was selected for its integration of ACT, an evidence-based treatment for depression and anxiety. The chatbot was developed in collaboration with clinical psychologists and researchers to integrate ACT-based principles. The chatbot includes features such as conversational AI, emotional logging, interactive exercises, gamification elements, and monitoring tools to support user engagement. This therapeutic approach enables the AI chatbot emol to offer targeted, structured interventions, distinguishing it from chatbots with less comprehensive therapeutic frameworks. The user interface uses a minimalist design, with simple text-based

interactions and no external hyperlinks. Content was developed in collaboration with clinical psychologists and researchers at Emol Inc, ensuring alignment with ACT-based principles. The chatbot provides asynchronous communication, allowing users to engage at their convenience without requiring real-time interaction.

The AI chatbot emol's design prioritizes accessibility and engagement, particularly for young users, by featuring a friendly AI character named Roku. The chatbot provides asynchronous communication, allowing participants to engage at their convenience. Sessions are preprogrammed to follow a consistent instructional strategy, beginning with a mood check-in and concluding with goal-setting exercises. Additional features, such as emotional logging and sleep tracking, allow users to actively monitor their mental health. The AI chatbot emol's comprehensive approach to digital mental self-care makes it especially well-suited for adolescents, who may prefer digital interactions over traditional therapy. By offering an approachable alternative for managing mental health challenges, emol has the potential to fill a critical gap in mental health service delivery for young people awaiting professional care.

However, no formal feasibility or usability studies have been conducted specifically for children and adolescents awaiting psychiatric consultation. The selection of emol for this study was based on a comprehensive review of Japanese mental health technology apps and direct discussions with the developers.

In this study, Emol Inc provided emol at a discounted rate. This is disclosed in the Conflicts of Interest section.

Details regarding the character design of emol and its ACT-based interactions with the character Roku are provided in Multimedia Appendix 1.

Participants

Participants in this study were individuals on the psychiatric waiting list for Yokohama City University Hospital, Yokohama City University Medical Center, Kanagawa Children's Medical Center, and Fujisawa City Hospital; all 4 hospitals are located in the southeastern region of Kanagawa Prefecture. These hospitals play a central role in child psychiatric services in the eastern part of Kanagawa Prefecture, Japan. Participants had to be aged between 12 and 18 years at the time of enrollment. As of January 2024, the population of 10‐ to 19-year-olds in Kanagawa Prefecture was estimated to be 774,283 [21].

The recruitment period was planned to span from October 1, 2023, to September 30, 2024, allowing sufficient time to ensure robust participant enrollment and data collection.

Inclusion and Exclusion Criteria

The inclusion criteria were as follows: (1) a score of 10 or higher on the PHQ-9, indicating clinical depressive symptoms; (2) access to a device capable of running emol; (3) access to online interviews via Zoom (Zoom Inc); (4) proficiency in reading and writing Japanese at the upper elementary school level; and (5) agreement to abstain from using other mental self-care apps during the study period.

The exclusion criteria were as follows: (1) patients deemed to require urgent care by a child psychiatrist, (2) patients reporting a suicide attempt within the past 2 weeks on their screening questionnaire or patients in urgent need of physical treatment due to conditions such as anorexia, (3) patients who had received treatment at another psychiatric facility within the past month or who expressed a desire for future treatment at such a facility, (4) patients who were continuing to take psychotropic medications prescribed by the above-mentioned facilities, and (5) patients unable to complete diaries due to physical issues such as injury.

Intervention Group

The intervention group received general mental health information via the Yokohama City University Child Psychiatry Department's website, "Oyako-no Kokoro-no Tomarigi" [22], which provides short video programs and texts that provide easy-to-understand explanations of common mental health issues for children and adolescents.

In addition, 8 weekly sessions were provided by emol. Each session lasted between 20 and 30 minutes. No major bug fixes, system downtimes, content changes, or unexpected events occurred during the trial that could have influenced the intervention's functionality or delivery. During the trial period, the version of emol remained unchanged, and no major content updates, feature modifications, or dynamic content changes occurred. The intervention was evaluated as a stable version, ensuring the replicability of study findings. While the current version reflects the intervention used in this study, future changes may occur. To ensure replicability, screenshots and videos of the interface are available upon request from the developers. Participants accessed emol via their personal smartphones. All participants needed to have a stable internet connection and a device capable of running the app. Access to the chatbot was restricted to study participants only, and no public demo mode was available. However, a free sample version of the emol program, which differs from the study version, is publicly available on the website. Participants were not required to pay any fees for accessing the chatbot during the study. A stable internet connection and a compatible device were necessary for access. No critical secular events, such as changes in internet resources or hardware requirements, occurred during the study period.

While we await specific dialogue examples from Emol Inc to further illustrate how these ACT processes were implemented conversationally, the session structure was designed to progressively build psychological flexibility through these interconnected processes. The detailed session structure and content are presented in Multimedia Appendix 2, and an example of a user interaction with the AI character Roku is illustrated in Multimedia Appendix 1.

The session overview in emol included the following:

- Session 1 (distress and enduring): introduces psychological flexibility by helping users recognize their distress patterns, aligning with the acceptance process of ACT
- Session 2 (avoidance of experiences): addresses cognitive defusion by identifying experiential avoidance patterns and

encouraging users to observe their thoughts rather than becoming entangled in them
- Session 3 (control over what can be managed): emphasizes being present and distinguishing between controllable and uncontrollable aspects of experience
- Session 4 (acceptance): deepens the acceptance process by guiding users to embrace difficult emotions rather than struggling with them
- Session 5 (observing from a detached perspective): develops self-as-context by helping users adopt an observer perspective toward their experiences
- Session 6 (life values): explores the values process by assisting users in identifying what matters most to them personally
- Session 7 (commitment): focuses on committed action by translating values into concrete behavioral goals
- Session 8 (continuation of commitment): reinforces committed action while integrating all ACT processes for sustainable change

Weekly online assessments were conducted at week 0, during the intervention period, and at week 9. No additional cointerventions, training sessions, or structured support were provided beyond these assessments. For routine application outside of the RCT setting, no training is required for users. No automated prompts or reminders were used during the trial or for routine application outside of the RCT setting. The periodic assessments were administered by nonphysician research assistants, who performed minimal mental status checks and checked assessment items, such as PHQ-9, Athens Insomnia Scale (AIS), and adverse events. Simultaneously, nonphysician research assistants saved recorded audio data for voice analysis. They also provided technical assistance when required. Human support was limited to assessment and monitoring during the intervention, with no direct involvement in the therapeutic process. For routine application outside of the RCT setting, no human involvement would be required. For measuring writing pressure, participants were provided with an intelligent pen pressure device developed by Zebra Holdings Inc and given a diary in which they recorded the date, time, weather, and mood (options included “good,” “normal,” or “bad”) within 2 hours of waking. Research assistants encouraged participants to use the pen consistently for their diary entries.

Control Group

The control group received general mental health information via the Yokohama City University child psychiatry department’s website, “Oyako-no Kokoro-no Tomarigi” [22]. This website provides educational resources about common mental health conditions in children and adolescents through easy-to-understand videos and text explanations specifically designed for young people. The video content features conversations between teddy bear and rabbit avatars discussing common mental health symptoms and concerns in children and adolescents, followed by child-friendly explanations from a child psychiatrist. Topics covered in these educational videos include suicidal thoughts, lack of energy/motivation, anxiety, isolation and loneliness, obsessive worrying, attention difficulties, self-harm behaviors, sleep problems, and auditory hallucinations. The child psychiatrist appearing in these videos is one of the authors of this study (JF). The website also contains separate sections with mental health resources for children and families, including multiple question and answer entries about children’s mental health issues. These materials are purely informational and educational in nature, rather than providing interactive or personalized therapeutic interventions. Participants were free to view the videos at their own discretion, without a predefined schedule. However, research assistants confirmed and recorded whether participants had viewed the assigned video content during each assessment session.

Unlike the intervention group, participants in the control group did not receive any structured therapeutic interaction through the AI chatbot. This design allowed for comparison between passive information provision (control) and active personalized therapeutic engagement (intervention).

Participants in the control group underwent the same regular online evaluations and diary recording process as those in the intervention group.

Criteria for Discontinuation

Criteria outlining conditions for discontinuation of individual participants and termination of the study are mentioned in Textbox 1.

Textbox 1. The criteria for discontinuing the intervention.

The criteria for discontinuing the intervention for individual participants were as follows:

1. Withdrawal of consent by the participant or their legal representative.
2. Determination that the participant no longer met the inclusion criteria or met any exclusion criteria postenrollment.
3. Worsening of symptoms or findings that made study continuation challenging.
4. The occurrence of adverse events that posed challenges to study continuation.
5. Initiation of additional treatments, such as psychiatric care, counseling, or the use of mental health apps.
6. A determination by the principal investigator or sub-investigator that continuation was otherwise undesirable.

The criteria for study termination were as follows:

1. Determination that the study intervention lacked expected efficacy, posed safety concerns, or was no longer meaningful to continue.
2. Significant delays in case registration, frequent protocol deviations, or other factors made the study completion difficult.
3. Occurrence of serious compliance issues affecting study execution.

Primary Outcome

The primary outcome was the change in PHQ-9 scores from pre- to postintervention. This measure was selected because assessing treatment efficacy through PHQ-9 score changes is commonly recommended and widely accepted in clinical research, including in Japan [23]. The PHQ-9, a self-administered questionnaire consisting of nine items, evaluates the presence and severity of depressive symptoms based on *Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition* (*DSM-IV*) criteria for major depressive disorder within the past 2 weeks. The total PHQ-9 score ranges from 0 to 27, with higher scores indicating more severe depressive symptoms. The PHQ-9 score obtained at week 0 was used as the baseline.

Secondary Outcomes

The secondary outcome measures evaluated the correlation and relationship between changes in the primary outcome measure (PHQ-9) and the following items.

Athens Insomnia Scale

The AIS is a self-assessment tool developed as part of the World Health Organization's "World Project on Sleep and Health" to evaluate insomnia with high reliability and validity [24]. The scale includes 8 items, with 5 assessing nighttime sleep difficulties ("sleep onset," "nighttime awakenings," "early morning awakenings," "sufficiency of total sleep duration," and "satisfaction with sleep quality") and 3 evaluating daytime functional impairment ("daytime mood," "daytime activity level," and "daytime sleepiness"). Participants rated the frequency of these experiences (at least 3 times a week in the past month) on a 4-point scale. A total score of 4 or more suggested suspected insomnia, while a score of 6 or more indicated insomnia.

Voice Analysis

During online consultations, participants' voices were recorded while they conversed with the interviewer. Voice data was analyzed by SHIN4NY Inc, with recordings provided to the company for analysis. A previous study demonstrated that a speech emotion recognition model could predict depression [25].

Writing Pressure Analysis

Upon providing consent, participants received a diary and instructions on how to record entries using an intelligent pen pressure device developed by ZEBRA HOLDINGS Inc. This device, a mechanical pencil, automatically captured data on writing pressure, acceleration, and pen angle, storing it without requiring additional action from the participant. Previous research has suggested that an intelligent pen capable of measuring writing pressure may predict anxiety levels [26].

These data were collected at baseline (week 0), during the intervention, and at the study's conclusion (week 9) through online consultations. Evaluations included PHQ-9, AIS, voice analysis, and intelligent pen pressure device data.

Data Management

Data entry was conducted by research assistants and verified by the principal investigator. A data dictionary guided the coding process, and central monitoring took place once during the study period based on the collected case report forms.

This study managed data for both clinical outcomes and user engagement. In addition to the primary and secondary outcome measures, engagement metrics were tracked using CSV data logs provided by Emol Inc, including (1) total usage time, (2) average daily usage time, (3) usage time periods, (4) last completed session, (5) last usage date, (6) first usage date, and (7) session progression history. The purpose of analyzing these engagement metrics was 2-fold. First, we aimed to identify usage patterns and their potential relationship with changes in depressive symptoms. Second, these data were intended to assess whether participants engaged appropriately with the app and to evaluate whether they met the expected level of engagement for effective intervention. For example, adequate engagement might be defined as completing a minimum number of chatbot sessions or maintaining consistent interaction over time.

Statistical Analysis

The analysis populations were defined as follows, with the primary analysis population being the full analysis set (FAS). This study included all participants who were registered, randomized, received at least one session of the trial intervention, and had available efficacy data. Participants were excluded if baseline data were not obtained or if significant protocol violations occurred. In addition, the FAS included participants who had no major deviations from the study protocol. Missing data were analyzed as observed without imputation, with the primary analysis performed on the FAS. As this study was exploratory in nature and lacked previous research in this specific population, no imputation for missing data was conducted.

The target sample size for the study was 60 participants (30 in each group). This calculation assumed an expected between-group mean difference in PHQ-9 change scores of 7, an SD of 10, a 2-sided significance level of .10, and a power of 0.80, using an independent 2-sample *t* test. The estimated sample size was 26 participants per group, and a target of 30 participants per group was set to allow for possible exclusions.

The primary analysis was conducted using the FAS. Changes in PHQ-9 scores from pre- to postintervention were presented as mean (SD). An analysis of covariance estimated the difference between groups, including the 90% CI and *P* value for the group comparison. A 2-sided test was used, with statistical significance determined at $P<.10$. The secondary analysis was conducted on the per-protocol set using the same summary and analytical methods as the primary analysis.

Protocol Version and Amendments

Protocol version 1 was initially approved on September 27, 2023, with version 2 subsequently approved on February 15, 2024.

Ethical Considerations

This study was registered with the Japan Registry of Clinical Trials under registration number jRCT1032230427. The full trial protocol is available on the Japan Registry of Clinical Trials website [27]. Approval for the study was granted by the Ethics Committee of Yokohama City University (approval F230907001). All procedures were conducted in accordance with the ethical standards outlined in the Declaration of Helsinki and the "Ethical Guidelines for Medical and Health Research Involving Human Participants." Written informed consent was obtained from both participants and their guardians before enrollment. When possible, assent from participants was also sought; if direct assent was difficult to obtain, informed assent was acquired as an alternative. If participants exhibited severe psychological distress or suicidal ideation during the study, the research team had a protocol in place to guide them to appropriate psychiatric services. Emergency contact information for crisis intervention was provided to all participants and their guardians. The study adhered to safety monitoring procedures to ensure participant well-being throughout the intervention. Participant privacy and confidentiality were strictly protected, and collected data were used solely for research purposes. All personal information was handled in accordance with Yokohama City University's privacy policy [28]. Findings from the study are to be disseminated through peer-reviewed journals and academic conferences.

Results

This study was conducted across 4 hospitals in Kanagawa Prefecture: Kanagawa Children's Medical Center, Fujisawa City Hospital, Yokohama City University Hospital, and Yokohama City University Medical Center, targeting pediatric and adolescent psychiatric patients awaiting consultation. Recruitment materials were distributed from October 2023 to June 2024 to patients on waiting lists at each hospital as follows:

- Kanagawa Children's Medical Center: 78 patients
- Fujisawa City Hospital: 8 patients
- Yokohama City University Hospital: 10 patients
- Yokohama City University Medical Center: 0 patients

A total of 96 patients received study invitations (78 from Kanagawa Children's Medical Center, 8 from Fujisawa City Hospital, and 10 from Yokohama City University Hospital). Of these, 8 patients expressed interest by contacting us for additional information. Out of those who expressed interest, 3 patients scheduled and completed the informed consent process, while the remaining 5 did not proceed with informed consent due to a lack of response when attempting to arrange a schedule.

Among the 3 patients who completed the informed consent process, 1 participant (a female adolescent) provided consent but subsequently withdrew from the study. The participant's family initially contacted the research team on the scheduled day of the first online session, stating:

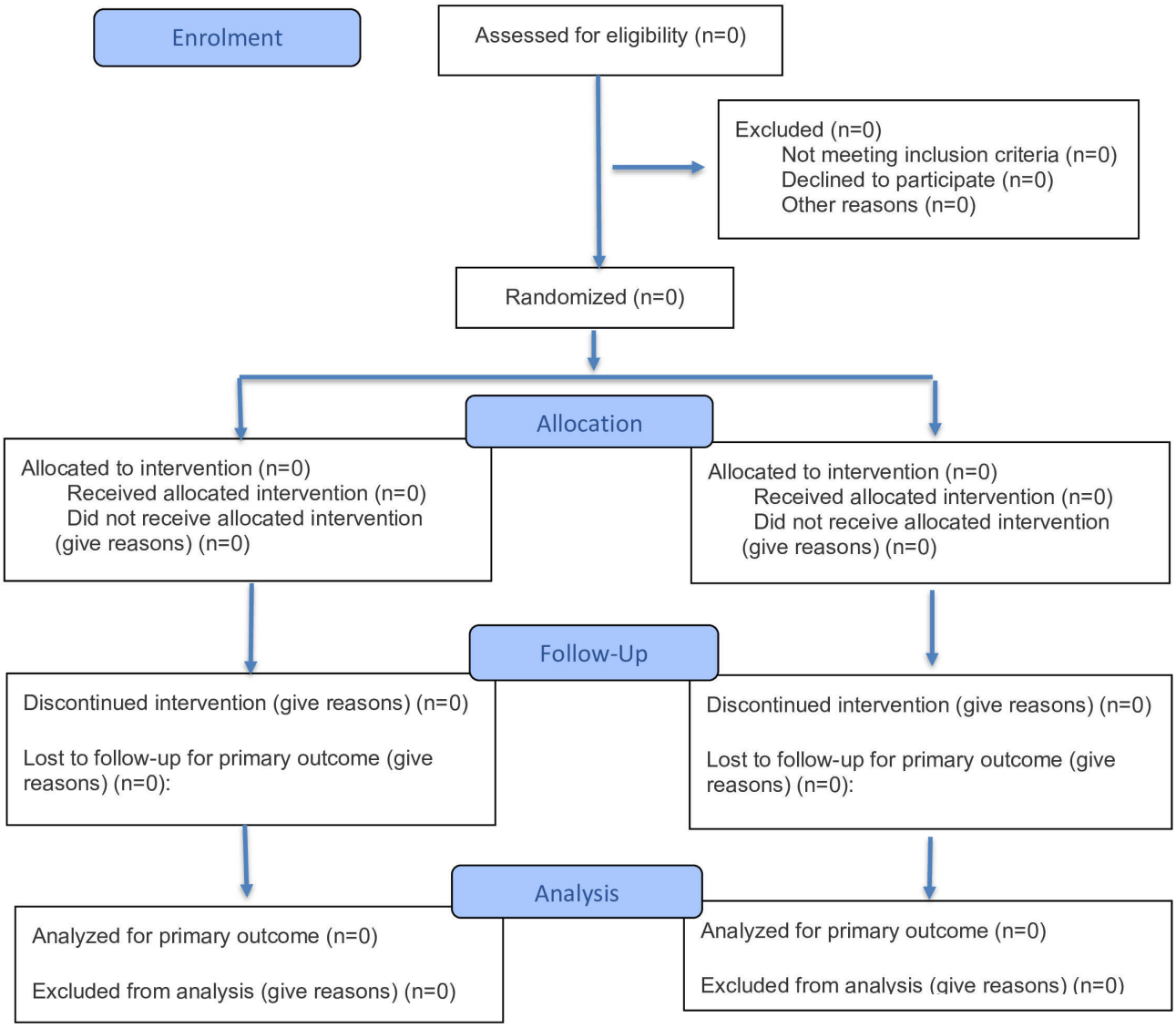
> *This morning, she became panic-stricken and is now unable to participate. Although it is the day of the appointment, would it be possible to cancel? I sincerely apologize for the inconvenience caused after all your preparations.*

In a follow-up message, the family elaborated:

> *She expressed anxiety about the online interview, making it impossible to proceed. We had hoped that engaging in this activity might help her develop a more positive outlook, but perhaps it was still too challenging for her.*

The other 2 patients who completed the informed consent process either declined participation due to concerns about diary recording requirements or were excluded after beginning medication at another facility. Another patient declined participation due to concerns about diary recording, and the third patient was excluded after beginning medication at another facility. Consequently, no evaluable data were obtained in this study (Figure 1).

Figure 1. CONSORT (Consolidated Standards of Reporting Trials) 2025 flow diagram.



Discussion

Barriers to Engagement in Digital Interventions

The primary finding of this study is that children and adolescents with mental health challenges and their caregivers showed limited interest in an AI chatbot-based intervention while awaiting psychiatric consultation, highlighting complex barriers to the adoption of digital mental health interventions. Several factors contributed to this low engagement. In this study, only one participant consented but subsequently withdrew due to a worsening of symptoms. This outcome suggests that patients with more severe symptoms may find digital interventions less suitable or accessible. However, given the need for timely intervention among high-severity cases, it is worth exploring ways to make AI-driven tools like chatbots more adaptable to their needs. Improving the usability and support level of digital tools for more severe cases could provide valuable early intervention options where in-person resources are limited.

Digital Intervention Challenges: Accessibility, Family Influence, and Psychological Burden

First, the one participant who withdrew provides valuable insights into digital intervention barriers. Despite digital tools being promoted as accessible for those with social anxiety, this case reveals that even virtual interactions can trigger a significant psychological burden in adolescents with severe mental health challenges. The mother's hope that "engaging in this activity might help her develop a more positive outlook" contrasted with her realization that "perhaps it was still too challenging for her," highlighting the gap between theoretical accessibility and practical barriers. The child psychiatrist who interviewed the patient and family during adverse event verification determined that the patient experienced increased subjective burden at the time of study participation, along with anxiety and avoidance symptoms exacerbated by her underlying depressive condition. This case demonstrates a "digital intervention paradox"—while technology aims to increase accessibility, implementation requirements like scheduled online sessions can create new barriers for the very individuals they intend to help. Future interventions might need truly

asynchronous options that minimize direct interaction while maintaining efficacy and safety monitoring.

Many young patients, particularly those who have not received a formal diagnosis or treatment, may struggle to understand the relevance of app-based mental health support and may not fully appreciate its therapeutic value, especially when compared to the immediate effects associated with in-person interventions like pharmacotherapy or face-to-face therapy. A previous study with small sample sizes has shown that AI chatbot app dropouts for psychogenic premature ejaculation were only around 25% among Japanese adults [16]. On the other hand, previous studies suggest that young patients often lack intrinsic motivation to engage with mental health apps, especially when their symptoms are severe enough to require professional care [29,30]. In particular, previous research indicates that individuals with more severe symptoms may benefit from more personalized support to enhance engagement and adherence to digital interventions [31]. Without a clear understanding of the intervention or trust in its benefits, these young patients may lack motivation to use the app consistently, contributing to low engagement in the study.

Study Protocol Complexity and Privacy Concerns

This study may have unintentionally targeted a population less receptive to alternative digital interventions. Families who had already secured an upcoming psychiatric appointment may have seen little value in participating in a study involving digital interventions, preferring instead to wait for their scheduled in-person consultation. For these families, traditional in-person care may have appeared more reassuring, especially given the severity of the patient's symptoms. Previous research on social influences in mental health service-seeking behavior among young people suggests that family is often the primary influence in choosing in-person services, whereas young people themselves tend to make decisions regarding online services [4]. Another study has also found that parents often seek informal support for their children's mental health concerns initially, only turning to professional services as issues become more severe [32]. In addition, patients with severe symptoms or their families often prefer in-person consultations over digital interventions, perceiving in-person care as more reliable and suitable for managing serious symptoms [33]. Therefore, patients and families may value the familiarity and perceived efficacy of traditional, in-person care as a more reliable or reassuring option compared to digital alternatives. This preference likely contributed to the reluctance toward digital solutions observed in this study. Engaging patients and families earlier in the mental health care process—before they have secured traditional clinical appointments—might improve receptiveness to digital options. While an active control group could have offered a more rigorous comparison, we selected a passive control condition due to practical constraints. At the time of study planning, emol was the only adolescent-appropriate AI chatbot in Japan that integrated evidence-based psychological content (ACT), had a suitable user interface, and was available for research use. No other comparable tool was identified. Thus, we chose a passive control to reflect the real-world conditions in Japan, where patients on psychiatric waiting lists typically receive only basic informational support.

Furthermore, our study protocol involved multiple evaluation sessions, which could have imposed additional stress on participants, particularly those with social anxiety or other issues related to interpersonal interactions. Previous studies indicate that such requirements can increase dropout rates, especially among adolescents with social phobia [34,35]. In this context, the protocol's demands may have discouraged participation and contributed to low engagement. Although efforts were made to reduce participant burden by conducting interviews online, the process of obtaining informed consent and providing detailed study information likely remained a challenge for some participants, especially those who may be sensitive to social interactions. Moreover, secondary assessment methods like voice and writing pressure analysis may have caused some participants to feel self-conscious or concerned about privacy, potentially exacerbating symptoms or causing reluctance to participate [36]. Privacy concerns are common with mental health apps, especially when personal data is analyzed, and this may affect user engagement [37]. In this study, participants may have felt uneasy about the voice recordings or writing pressure data being analyzed, as well as discussing their mental health status with researchers. These issues may have created additional barriers to engagement, particularly for young people unfamiliar with research environments. In the current study design, the exclusion of severe cases and the online interview assessment to reduce the burden of face-to-face implementation may not have been sufficient.

Strengths and Limitations

This study had both methodological strengths and limitations. Conducting an RCT in an understudied population—children and adolescents on psychiatric waiting lists—provided valuable insights into a critical phase of mental health service delivery, and the exploratory nature of the trial allowed for close examination of data acquisition and dropout rates. In addition, our use of multiple assessment methods enabled a comprehensive evaluation of potential therapeutic effects and engagement patterns, which is rare in digital mental health studies involving youth.

However, the intensity of the intervention, including frequent evaluations and a structured protocol, may have been overwhelming for participants. This study suggests that less rigid protocols with fewer demands could enhance engagement. In addition, recruiting patients already awaiting traditional psychiatric care may have reduced openness to digital alternatives, thus limiting the generalizability of our findings. Targeting a population that has not yet committed to in-person care—such as students in school counseling or those referred by community organizations—may address this issue in future studies. Finally, while secondary measures such as voice and writing pressure analysis provide valuable data, they may also create privacy concerns that deter participation, particularly among younger users. In addition, typical limitations of eHealth trials should be noted. Participants were not blinded, which may have introduced performance bias. The informed consent process could have influenced their expectations, and the multiple outcomes planned for the study could have increased the risk of type 1 errors. Future studies should address these biases and explore strategies to improve participant engagement.

Furthermore, a key limitation of this study was the lack of a structured plan for conducting qualitative interviews and analyses on dropout reasons. This was primarily because the initial study design focused on quantitative outcome measures, and the feasibility of integrating additional qualitative assessments was not fully considered. While we identified some possible factors, such as social anxiety, depressive symptoms, and difficulties with maintaining engagement, a more systematic approach to understanding participants' experiences and challenges would have provided deeper insights.

Future studies should incorporate structured qualitative methods, such as exit interviews or surveys, to better understand engagement barriers and develop targeted strategies for improving retention and adherence in digital mental health interventions for adolescents on psychiatric waiting lists. Previous qualitative studies have identified key factors influencing user engagement, including personalization, trust in AI, and perceived relevance of content [38]. Integrating these insights into future chatbot interventions may enhance usability and acceptability.

Conclusions

To better understand the potential of AI chatbot interventions like emol, future studies could test the app in environments where supportive, pre-existing relationships exist, such as schools or community youth centers. Conducting trials in these familiar settings may foster trust, encourage participation, and enhance data validity, ultimately increasing the accessibility and effectiveness of digital mental health interventions for younger adults.

Acknowledgments

We would like to express our gratitude to the following individuals for their invaluable support in this study. Minori Ito assisted with participant recruitment and study progress management as a research assistant. Mikiko Yuzawa from Nouvelle Plus Inc. managed the data center and prepared the ethics application documents. Yasuyuki Okumura from the Initiative for Clinical Epidemiological Research provided support in developing the research plan. We also thank Daiki Takekawa from Emol Inc, Eiji Okuno from ZEBRA HOLDINGS Inc, and Kanji Okazaki from SHIN4NY Inc for their contributions in providing devices and support from their companies. We also acknowledge the assistance of OpenAI's ChatGPT and Anthropic's Claude in supporting the drafting and English editing process of this manuscript.

This study was funded by the Japan Science and Technology Agency under the Co-creation Opportunity Formation Support Program (FY2022–2031).

Data Availability

No datasets were generated or analyzed during this study.

Authors' Contributions

JF was responsible for conceptualization, methodology, investigation, writing of the original draft, supervision, project administration, and funding acquisition. YY contributed to methodology and writing—review and editing. SS carried out formal analysis and methodology, and contributed to writing—review and editing. NS participated in investigation and writing—review and editing. MO contributed to investigation and writing—review and editing. AS was involved in investigation, supervision, and writing—review and editing. MT and T Moroga contributed to investigation and data curation. HY and MI provided supervision and writing—review and editing. T Miyazaki contributed resources and writing—review and editing.

Conflicts of Interest

SS, NS, MO, MI, MT, HY, MT, and T Miyazaki have no conflict of interest. Individually, JF received research grants from KAKENHI (21K01994). JF also served as a member of an advisory board for Seisa Yokohama Educational Counseling Center. YY received consulting fees from ZEBRA HOLDINGS Inc. In this study, Emol Inc provided the AI Chatbot emol at a discounted rate, ZEBRA HOLDINGS Inc contributed research funding, and SHIN4NY Inc offered devices at a reduced cost. The supporting companies and the funding agency had no role in study design, data collection, analysis, interpretation, or manuscript submission decisions.

Multimedia Appendix 1
Screenshots of the artificial intelligence chatbot emol interface.
[PNG File, 60 KB - xmed_v6i1e70960_app1.png]

Multimedia Appendix 2
Session structure of the AI chatbot emol aligned with Acceptance and Commitment Therapy core processes.
[DOCX File, 17 KB - xmed_v6i1e70960_app2.docx]

Checklist 1
CONSORT-EHEALTH checklist (V 1.6.1).
[PDF File, 14863 KB - xmed_v6i1e70960_app3.pdf]

References

1. de Lijster JM, Dieleman GC, Utens E, et al. Social and academic functioning in adolescents with anxiety disorders: a systematic review. J Affect Disord 2018 Apr 1;230:108-117. [doi: 10.1016/j.jad.2018.01.008] [Medline: 29407534]
2. Verboom CE, Sijtsema JJ, Verhulst FC, Penninx B, Ormel J. Longitudinal associations between depressive problems, academic performance, and social functioning in adolescent boys and girls. Dev Psychol 2014 Jan;50(1):247-257. [doi: 10.1037/a0032547] [Medline: 23566082]
3. Davey CG, McGorry PD. Early intervention for depression in young people: a blind spot in mental health care. Lancet Psychiatry 2019 Mar;6(3):267-272. [doi: 10.1016/S2215-0366(18)30292-X] [Medline: 30502077]
4. Rickwood DJ, Mazzer KR, Telford NR. Social influences on seeking help from mental health services, in-person and online, during adolescence and young adulthood. BMC Psychiatry 2015 Mar 7;15:40. [doi: 10.1186/s12888-015-0429-6] [Medline: 25886609]
5. Kaku SM, Sibeoni J, Basheer S, et al. Global child and adolescent mental health perspectives: bringing change locally, while thinking globally. Child Adolesc Psychiatry Ment Health 2022 Nov 7;16(1):82. [doi: 10.1186/s13034-022-00512-8] [Medline: 36345001]
6. Sakano M, Snowden N. Paving the way for the future of child and adolescent mental health in Japan. London J Prim Care (Abingdon) 2018 Jul 4;10(4):123-125. [doi: 10.1080/17571472.2018.1483002]
7. Racine N, McArthur BA, Cooke JE, Eirich R, Zhu J, Madigan S. Global prevalence of depressive and anxiety symptoms in children and adolescents dring COVID-19: a meta-analysis. JAMA Pediatr 2021 Nov 1;175(11):1142-1150. [doi: 10.1001/jamapediatrics.2021.2482] [Medline: 34369987]
8. Hawton K, Saunders KEA, O'Connor RC. Self-harm and suicide in adolescents. Lancet 2012 Jun 23;379(9834):2373-2382. [doi: 10.1016/S0140-6736(12)60322-5] [Medline: 22726518]
9. Correll CU, Galling B, Pawar A, et al. Comparison of early intervention services vs treatment as usual for early-phase psychosis: a systematic review, meta-analysis, and meta-regression. JAMA Psychiatry 2018 Jun 1;75(6):555-565. [doi: 10.1001/jamapsychiatry.2018.0623] [Medline: 29800949]
10. Nakao M, Shirotsuki K, Sugaya N. Cognitive–behavioral therapy for management of mental health and stress-related disorders: recent advances in techniques and technologies. BioPsychoSocial Med 2021 Dec;15(1):16. [doi: 10.1186/s13030-021-00219-w]
11. Andrews G, Basu A, Cuijpers P, et al. Computer therapy for the anxiety and depression disorders is effective, acceptable and practical health care: an updated meta-analysis. J Anxiety Disord 2018 Apr;55(2):70-78. [doi: 10.1016/j.janxdis.2018.01.001]
12. Fitzpatrick KK, Darcy A, Vierhile M. Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (Woebot): a randomized controlled trial. JMIR Ment Health 2017 Jun 6;4(2):e19. [doi: 10.2196/mental.7785] [Medline: 28588005]
13. Fulmer R, Joerin A, Gentile B, Lakerink L, Rauws M. Using psychological artificial intelligence (Tess) to relieve symptoms of depression and anxiety: randomized controlled trial. JMIR Ment Health 2018 Dec 13;5(4):e64. [doi: 10.2196/mental.9782] [Medline: 30545815]
14. Zhong W, Luo J, Zhang H. The therapeutic effectiveness of artificial intelligence-based chatbots in alleviation of depressive and anxiety symptoms in short-course treatments: a systematic review and meta-analysis. J Affect Disord 2024 Jul 1;356:459-469. [doi: 10.1016/j.jad.2024.04.057] [Medline: 38631422]
15. Kamita T, Matsumoto A, Ito T, Inoue T. Development and evaluation of a chatbot system for self-guided mental healthcare based on the SAT counseling method [article in japanese]. Information Processing Society of Japan. 2020 May 27. URL: https://ipsj.ixsq.nii.ac.jp/record/204708/files/IPSJ-DCC20025004.pdf [accessed 2025-09-04]
16. Saito J, Kumano H, Ghazizadeh M, Shimokawa C, Tanemura H. An acceptance and commitment therapy smartphone application for erectile dysfunction: a feasibility study. Curr Ther Res Clin Exp 2023;99:100728. [doi: 10.1016/j.curtheres.2023.100728] [Medline: 38090722]
17. Cao J, Kurata K, Lim Y, Sengoku S, Kodama K. Social acceptance of mobile health among young adults in Japan: an extension of the UTAUT model. Int J Environ Res Public Health 2022 Nov 17;19(22):15156. [doi: 10.3390/ijerph192215156] [Medline: 36429875]
18. Eleuch AEK. Healthcare service quality perception in Japan. Int J Health Care Qual Assur 2011;24(6):417-429. [doi: 10.1108/09526861111150680] [Medline: 21916144]
19. Shimizu Y, Ishizuna A, Osaki S, et al. The social acceptance of smart health services in Japan. Int J Environ Res Public Health 2022 Jan 24;19(3):1298. [doi: 10.3390/ijerph19031298] [Medline: 35162321]
20. Emol. URL: https://emol.jp/ [accessed 2025-08-28]
21. Kanagawa prefecture age-specific demographic survey results [Web Page in Japanese]. Kanagawa Prefecture. URL: https://www.pref.kanagawa.jp/docs/x6z/tc30/jinko/nenreibetu.html [accessed 2025-08-28]

22. Notice from the Child Psychiatry Department [Web Page in Japanese]. Yokohama City University Hospital. URL: https://www-user.yokohama-cu.ac.jp/~ycucap/ [accessed 2025-08-28]
23. Muramatsu K, Miyaoka H, Kamijima K, et al. Performance of the Japanese version of the Patient Health Questionnaire-9 (J-PHQ-9) for depression in primary care. Gen Hosp Psychiatry 2018;52:64-69. [doi: 10.1016/j.genhosppsych.2018.03.007] [Medline: 29698880]
24. Okajima I, Nakajima S, Kobayashi M, Inoue Y. Development and validation of the Japanese version of the Athens Insomnia Scale. Psychiatry Clin Neurosci 2013 Sep;67(6):420-425. [doi: 10.1111/pcn.12073] [Medline: 23910517]
25. Hansen L, Zhang YP, Wolf D, Sechidis K, Ladegaard N, Fusaroli R. A generalizable speech emotion recognition model reveals depression and remission. Acta Psychiatr Scand 2022 Feb;145(2):186-199. [doi: 10.1111/acps.13388] [Medline: 34850386]
26. Tapia-Jaya C, Ojeda-Zamalloa I, Robles-Bykbaev V, Pesántez-Avilés F, San Andrés Becerra I, et al. An intelligent pen to assess anxiety levels through pressure sensors and fuzzy logic. In: Advances in Human Factors in Wearable Technologies and Game Design: Proceedings of the AHFE 2017 International Conference on Advances in Human Factors and Wearable Technologies, July 17-21, 2017, The Westin Bonaventure Hotel, Los Angeles, California, USA: Springer; 2017:64-71. [doi: 10.1007/978-3-319-60639-2_7]
27. A pilot study for an intervention study using AI chatbot-based cognitive behavioural therapy for depressed child and adolescent in waiting list. (None). Japan Registry of Clincal Trials. 2023 Oct 31. URL: https://jrct.niph.go.jp/re/reports/detail/86750 [accessed 2025-08-28]
28. Privacy policy [Web Page in Japanese]. Yokohama City University. URL: https://www.yokohama-cu.ac.jp/policy/privacy.html [accessed 2025-08-28]
29. Garrido S, Oliver E, Chmiel A, Doran B, Boydell K. Encouraging help-seeking and engagement in a mental health app: what young people want. Front Digit Health 2022;4:1045765. [doi: 10.3389/fdgth.2022.1045765] [Medline: 36620186]
30. Jebbink M. The motives of young adults to make use of mental health related apps in their daily lives -- a qualitative interview study [Master's thesis]. : University of Twente; 2021 Feb 26.
31. Götzl C, Hiller S, Rauschenberg C, et al. Artificial intelligence-informed mobile mental health apps for young people: a mixed-methods approach on users' and stakeholders' perspectives. Child Adolesc Psychiatry Ment Health 2022 Nov 17;16(1):86. [doi: 10.1186/s13034-022-00522-6] [Medline: 36397097]
32. Sawrikar V, Van Dyke C, Smith Slep AM. The Ws of parental help-seeking: when, where, and for what do parents seek help for child mental health. Child Psychiatry Hum Dev 2024 Mar 20. [doi: 10.1007/s10578-024-01683-5] [Medline: 38507021]
33. Apolinário-Hagen J, Harrer M, Kählke F, Fritsche L, Salewski C, Ebert DD. Public attitudes toward guided internet-based therapies: web-based survey study. JMIR Ment Health 2018 May 15;5(2):e10735. [doi: 10.2196/10735] [Medline: 29764797]
34. Christensen H, Griffiths KM, Farrer L. Adherence in internet interventions for anxiety and depression. J Med Internet Res 2009 Apr 24;11(2):e13. [doi: 10.2196/jmir.1194] [Medline: 19403466]
35. Melville KM, Casey LM, Kavanagh DJ. Dropout from internet-based treatment for psychological disorders. Br J Clin Psychol 2010 Nov;49(Pt 4):455-471. [doi: 10.1348/014466509X472138] [Medline: 19799804]
36. Gulliver A, Griffiths KM, Christensen H. Perceived barriers and facilitators to mental health help-seeking in young people: a systematic review. BMC Psychiatry 2010 Dec 30;10:113. [doi: 10.1186/1471-244X-10-113] [Medline: 21192795]
37. Radovic A, Gmelin T, Stein BD, Miller E. Depressed adolescents' positive and negative use of social media. J Adolesc 2017 Feb;55:5-15. [doi: 10.1016/j.adolescence.2016.12.002] [Medline: 27997851]
38. Saadati SA, Saadati SM. The role of chatbots in mental health interventions: user experiences. AI Tech Behav Soc Sci 2023;1(2):19-25. [doi: 10.61838/kman.aitech.1.2.4]

Abbreviations

ACT: acceptance and commitment therapy
AI: artificial intelligence
AIS: Athens Insomnia Scale
CBT: cognitive behavioral therapy
***DSM-IV*:** *Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition*
FAS: full analysis set
PHQ-9: 9-item Patient Health Questionnaire
RCT: randomized controlled trial

Edited by S Amal; submitted 07.01.25; peer-reviewed by B Tosti, E Ennis, MDG Ambrosio; revised version received 14.06.25; accepted 27.07.25; published 05.09.25.

Please cite as:
Fujita J, Yano Y, Shinoda S, Sho N, Otsuki M, Suda A, Takayama M, Moroga T, Yamaguchi H, Ishii M, Miyazaki T
Challenges in Implementing a Mobile AI Chatbot Intervention for Depression Among Youth on Psychiatric Waiting Lists: Randomized Controlled Study Termination Report
JMIRx Med 2025;6:e70960
URL: https://xmed.jmir.org/2025/1/e70960
doi:10.2196/70960

Converting Organic Municipal Solid Waste Into Volatile Fatty Acids and Biogas: Experimental Pilot and Batch Studies With Statistical Analysis

Hojjat Borhany, MSc

Faculty of Environmental Science, Department of Environmental Science, Informatic, and Statistics, University of Ca' Foscari Venice, Mestre (VE), Italy

Corresponding Author:
Hojjat Borhany, MSc
Faculty of Environmental Science, Department of Environmental Science, Informatic, and Statistics, University of Ca' Foscari Venice, Mestre (VE), Italy

Related Articles:

Companion article: https://www.biorxiv.org/content/10.1101/2023.06.21.545938v2

Companion article: https://med.jmirx.org/2025/1/e69895

Companion article: https://med.jmirx.org/2025/1/e69896

Companion article: https://med.jmirx.org/2025/1/e69894

Abstract

Background: Italy can augment its profit from biorefinery products by altering the operation of digesters or different designs to obtain more precious bioproducts like volatile fatty acids (VFAs) than biogas from organic municipal solid waste. In this context, recognizing the process stability and outputs through operational interventions and its technical and economic feasibility is a critical issue. Hence, this study involves an anaerobic digester in Treviso in northern Italy.

Objective: This research compares a novel line, consisting of pretreatment, acidogenic fermentation, and anaerobic digestion, with single-step anaerobic digestion regarding financial profit and surplus energy. Therefore, a mass flow model was created and refined based on the outputs from the experimental and numerical studies. These studies examine the influence of hydraulic retention time (HRT), pretreatment, biochar addition, and fine-tuned feedstock/inoculum (FS/IN) ratio on bioproducts and operational parameters.

Methods: VFA concentration, VFA weight ratio distribution, and biogas yield were quantified by gas chromatography. A *t* test was then conducted to analyze the significance of dissimilar HRTs in changing the VFA content. Further, a feasible biochar dosage was identified for an assumed FS/IN ratio with an adequately long HRT using the first-order rate model. Accordingly, the parameters for a mass flow model were adopted for 70,000 population equivalents to determine the payback period and surplus energy for two scenarios. We also explored the effectiveness of amendments in improving the process kinetics.

Results: Both HRTs were identical concerning the ratio of VFA/soluble chemical oxygen demand (0.88 kg/kg) and VFA weight ratio distribution: mainly, acetic acid (40%), butyric acid (24%), and caproic acid (17%). However, a significantly higher mean VFA content was confirmed for an HRT of 4.5 days than the quantity for an HRT of 3 days (30.77, SD 2.82 vs 27.66, SD 2.45 g–soluble chemical oxygen demand/L), using a *t* test (t_8=−2.68; *P*=.03; CI=95%). In this research, 83% of the fermented volatile solids were converted into biogas to obtain a specific methane (CH_4) production of 0.133 CH_4-Nm^3/kg–volatile solids. While biochar addition improved only the maximum methane content by 20% (86% volumetric basis [v/v]), the FS/IN ratio of 0.3 volatile solid basis with thermal plus fermentative pretreatment improved the hydrolysis rate substantially (0.57 vs 0.07, 1/d). Furthermore, the biochar dosage of 0.12 g-biochar/g–volatile solids with an HRT of 20 days was identified as a feasible solution. Principally, the payback period for our novel line would be almost 2 years with surplus energy of 2251 megajoules [MJ] per day compared to 45 years and 21,567 MJ per day for single-step anaerobic digestion.

Conclusions: This research elaborates on the advantage of the refined novel line over the single-step anaerobic digestion and confirms its financial and technical feasibility. Further, changing the HRT and other amendments significantly raised the VFA concentration and the process kinetics and stability.

(JMIRx Med 2025;6:e50458) doi:10.2196/50458

KEYWORDS

multistep fermentation; specific methane production; anaerobic digestion; kinetics study; biochar; first-order; modified Gompertz; mass balance; waste management; environment sustainability

Introduction

The European Union annually generated about 110 million tons of organic waste in 2006, which excluded slurry and manure. This waste mainly came from the food industry (33%), agriculture and hunting (30%), and households (20%) [1]. Current Italian legislation forbids landfilling organic waste and requires treating it through biological and thermal processes like anaerobic digestion, composting, and incineration with high disposal costs for secondary waste flux (€75 - €125 per ton; a currency exchange rate of €1=US $1.05 is applicable) [2]. Under the pressure of exhaustible natural exploitation and increasing organic waste, the European Commission approved the circular economy action plan to promote sustainable recovery methods to reduce the secondary waste flux. The techniques recommended in the circular economy context assume a "take-use-reuse" viewpoint. Such an approach wants to close the circuit of cycles, extend product life, and treat the wastes as precious recyclable materials [3,4]. In this respect, the European Union states have deployed biological processes such as anaerobic digestion to gain either platform chemicals like volatile fatty acids (VFAs) or biogas from organic wastes produced in urban areas [5-9]. These products are extremely valuable in the era of environmental disasters, which have several consequences (eg, climate change), since they are renewable, sustainable, carbon-neutral, and compatible with current fossil-based fuel infrastructures [10].

Recent studies have aimed at finding a sequential reclaiming route to obtain various bioproducts such as VFAs and biohydrogen with a higher added-value market than bio-methane at distinct steps to either redesign the existing plants or integrate them into biorefinery platforms [11,12]. Various biological processes can convert different feedstock (eg, edible sugary crops, oil-bearing crops, livestock, waste sludge [WS], and food waste) into a range of biofuels, including bioethanol, biodiesel, bio-methane, and biohydrogen [10,13,14]. Biofuel production from edible crops is quite controversial in terms of food supply, ethical quandary, and insecure supply chain. However, food waste, WS, and livestock are omnipresent in urban and rural areas without widespread deployment in a biorefinery scheme. Accordingly, this research aims to convert organic municipal solid waste (OMSW), mainly from food waste, into VFAs and biogas.

This study examines the biological recovery route for OMSW for potential beneficial bioproducts and technical feasibility. This effort includes three steps: pretreatment, mesophilic acidogenic fermentation, and anaerobic digestion. Specifically, we endeavor to conceive how to make the process more profitable and practicable through operational amendments that change the share of methanogenesis and acidogenic routes in the final products (VFAs and biogas) [9] and lower the costs of the process in terms of energy and water consumption. Hence, determining a reasonably priced process with a desirable VFA-rich stream from acidogenic fermentation and a high methane (CH_4) yield from methanogenesis [15] could ultimately encourage full-scale commercialization. VFAs typically serve as platform chemicals for many processes (eg, biopolymer synthesis of polyhydroxyalkanoates [PHAs] [16-19]), which could be later recovered through biological processes to close the material life cycle.

The major bottleneck in anaerobic digestion of biowaste is at the hydrolysis step. Such a problem could be relieved by various methods such as pretreatment, an optimized feedstock/inoculum (FS/IN) ratio, and carbonaceous material addition, including biochar [20-22]. The latter method was recently realized to have numerous benefits to the process, such as improving the process stability, acceleration of the process rate, buffering potency and alkalinity, inhibitors adsorption, enriched microbial functionality, and electron transfer mechanism. As a result, it could improve CH_4 generation by fostering hydrolysis, acetogenesis, and methanogenesis [23]. The residual solids out of the multistep line of pretreatment followed by acidogenic fermentation plus anaerobic digestion can be used in a pyrolysis line for biochar and biofuel production to further lower the secondary waste flux [24]. This strategy provides several benefits, such as combating climate change and global soil degradation and addressing the rising energy demand.

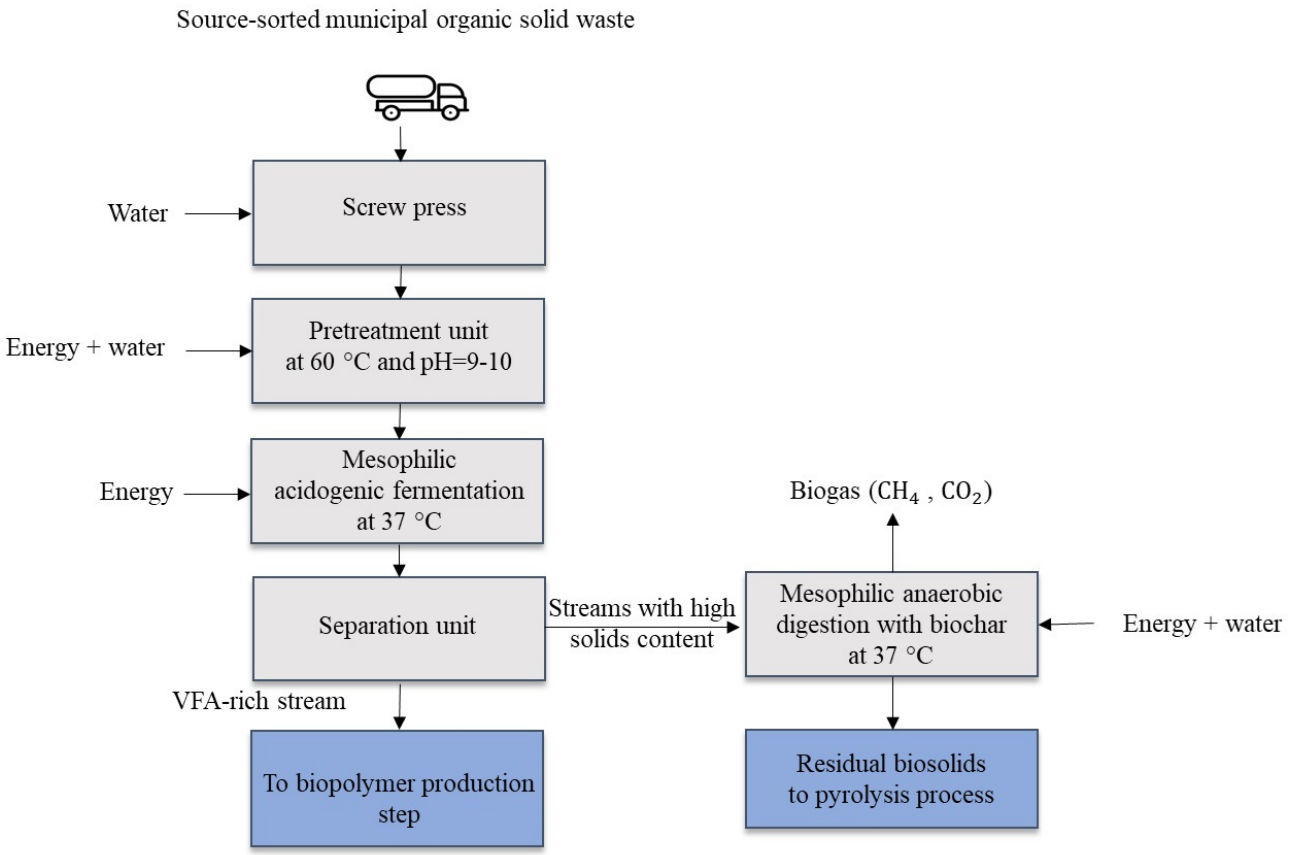
This study compares the multistep route of pretreatment, acidogenic fermentation, and anaerobic digestion with the existing method of single-step anaerobic digestion for valorizing OMSW in the Treviso wastewater treatment plant (WWTP) in terms of financial profit and technical feasibility. In this context, the present research has the ultimate goals of facilitating the entrance of the process into the market and further closure of the cycle of organic material. Accordingly, it assesses several suggestions, such as hydraulic retention time (HRT) variation, pretreatment, biochar addition, and adjusted FS/IN ratio to enhance the bioproducts and decrease the involved costs. To this end, their effects on the process were quantified through experimental tests, confirming their significance through statistical analysis. Later, the payback period, amount of surplus energy, and volatile solids (VS) destruction for the mentioned scenarios were determined using a mass balance model refined according to the laboratory studies. The boundary condition parameters for energy conversion and costs were assumed according to previous studies and experts' knowledge, respectively. To the best of the author's knowledge, this paper is novel in presenting a robust framework to assess a groundbreaking proposition for the valorization of OMSW financially and technically. Overall, we concluded that our line is viable technically and overtakes the conventional methods financially.

Methods

Biorefinery Process Scheme and Experimental Studies

Figure 1 presents the hypothesized biorefinery process line in this research. It comprises screw-pressing, a pretreatment unit, mesophilic acidogenic fermentation, solid-liquid separator, and mesophilic anaerobic digestion. The two sectors of biopolymer production and pyrolysis were exhibited differently since no mass and energy flow was considered for them, and only the possible end goals for the secondary stream were shown.

Figure 1. Schematic of the multistep process of pretreatment, acidogenic fermentation, and anaerobic digestion for VFAs and biogas production from the organic municipal solid waste. CH_4: methane; CO_2: carbon dioxide; VFA: volatile fatty acid.



After and before the pretreatment, the feedstock for different parameters was characterized from time to time. These parameters include the total solids (TS), VS, chemical oxygen demand (COD), soluble COD (SCOD), total Kjeldahl nitrogen, total phosphorous (P), ammonium ($N\text{-}NH_4^+$), phosphate ($P\text{-}PO_4^{3-}$), and VFA.

The feedstock that arrived at the WWTP had already been mixed with the acidogenic fermentative inoculum, which initiated solubilizing and converting the organic solid matters into SCOD and VFAs in the transporter. Then, in the pretreatment unit, a sodium hydroxide (NaOH) solution (40% kg/kg) was added to bring the pH to 9 - 10 and heated to 60 ˚C for 24 hours. Subsequently, the biomixture was fed manually into a 5 L (operational volume of 4.5 L) continuously stirred pilot acidogenic fermenter operated at the given conditions (Table 1). Its high alkalinity maintained the pH during the acidogenic fermentation in the optimal range. Further, the mixture was blended mechanically, and the whole system was kept in the oven to hold the temperature constant at 37 ˚C. The output was sampled frequently during the week, and the samples were centrifuged to obtain the supernatant to measure pH, SCOD, VFA, $N\text{-}NH_4^+$, and $P\text{-}PO_4^{3-}$. A tiny fraction of the residual solid part was used to characterize solids like COD, P, and total Kjeldahl nitrogen, and the rest was kept in the freezer to apply the bio-methane potential (BMP) test.

Table . The operational parameters of the mesophilic acidogenic fermenter.

Hydraulic retention time (days)	Organic loading rate (kg–volatile solids/m^3.d)	Temperature (˚C)	pH[a], mean (SD)
4.5	6.89	37	6.56 (0.25)
3	10.33	37	6.7 (0.45)

[a]13 measurements for pH.

The VS and TS characterization were performed in 105 ˚C and 550 ˚C ovens for 24 hours, respectively. Except for VFAs, all the remaining analyses (including COD measurements) followed the standard methods for examining water and wastewater [25]. The methods described in the A and D sections of No. 5220 for COD quantification were used. These methods are named "Closed Reflux, Titrimetric Method" and "Closed Reflux, Colorimetric Method" for the solid and liquid phases, respectively [25]. For the liquid, the samples were filtered after being centrifuged at 4500 rounds per minute (rpm) for 5 minutes, and before the analysis, the supernatant was filtered with a 0.45 µm cellulose filter (Whatman). For the solid, acidic digestion was performed at 220 ˚C with a high pressure of 2 atmospheres to destroy the 0.2 g of solid matrix for 2 hours. Afterward, the COD was measured in the solution using titration by ferrous ammonium sulfate as described in the standard methods. Our limit of detection was 50 - 500 mg-COD/L for the calorimetric method and 40 - 400 mg-COD/L for the titrimetric method. In this research, dilution was done for high-concentration values that are beyond the considered limit of detection.

In the BMP test, the effect of biochar addition was observed for 3 diverse dosages (0, 0.12, and 0.24 g-biochar/g-VS) on the bio-methane volume, content, and production kinetics in the mesophilic condition using four sets of the BMP test. The tests were conducted with a total number of 8 bottles of 250 mL (working volume of 215 mL). The anaerobic condition was ensured in bottles by sealing them after filling without any flushing with nitrogen molecules (N_2) or carbon dioxide (CO_2) since we had known that oxygen transfer at the surface of the waste stream was impossible as it contained a high TS and SCOD. This type of procedure was adopted in our laboratory and has been conducted for years. The biochar was synthesized by a local supplier, and its main physical and chemical features are reported in Table S1 in Multimedia Appendix 1. It was ground into microparticles and kept under a dried condition at room temperature before being added to the bottles. Further, the inoculum for the BMP test was collected from the 2300 m^3 completely stirred anaerobic digester treating thickened WS and squeezed OMSW mixture under the mesophilic condition at an organic loading rate (OLR) of 1.8 - 2.0 kg-VS/m^3.d in the treatment plant. The inoculum was added to the feedstock (residual solid from acidogenic fermentation) based on the weight ratio of 0.3 FS g-VS/IN g-VS. The TS and VS contents in the bottles (ie, inoculum and feedstock) were 133 g/kg and 17.6 g/kg, respectively.

The experiments were conducted for each condition, namely, only inoculum and either with or without biochar, in 2 bottles. The test was terminated after 25 days when the cumulative biogas production reached almost 89% of the final projected value. The biogas content was characterized by gas chromatography (for days 1, 4, 6, 10, 14, 16, 18, 21, and 25). Additionally, the values for the remaining days were filled through imputation using the *k*-nearest neighbors algorithm (number of neighbors=4 and weights=distance) [26]. The imputation code is provided in the repository [27]. Then, the biogas and bio-methane volumes were subtracted from the only inoculum to correct for the endogenous methane production, and both values were averaged for 2 bottles. Gas chromatography was performed using Agilent Technology (TM 6890N) with an HP-PLOT MoleSieve column (30 m length, 0.53 mm ID × 25 mm film thickness) and a thermal conductivity detector with argon as a carrier (79 mL/min). The hydrogen molecule (H_2), CH_4, oxygen molecule (O_2), and N_2 were analyzed using a thermal conductivity detector at 250 ˚C. The inlet temperature was 120 ˚C, with constant pressure in the injection port (ie, 70 kilopascal [kPa]). Samples were taken using a gas-type syringe (200 µL). Once the entire sample was vaporized, peak separation occurred within the column at a constant temperature of 40 ˚C for 8 minutes. We did not plan to monitor pH and other parameters like alkalinity, VFA, ammonia, and phosphate because the pH drop risk was negligible, and the biochar addition could provide a buffer capacity and adsorption of inhibitory compounds in the solution [28]. Moreover, a considerable part of the readily biodegradable COD of the feedstock was already converted to VFAs in the previous step. As a result, the process was easily controlled even in the transient condition when the risk of methanogenic inhibition was high [29].

Statistical Analysis and Performance Indicators

The performance indicators, including COD solubilization, VFA yield, ammonia and phosphate release, and VFA/SCOD ratio were determined. These indicators characterize the mesophilic acidogenic fermentation on the days when the data were available, and the process reached the pseudo-steady state condition. The indicators were calculated, and the data were plotted using a Microsoft Excel spreadsheet (Version 2412). In addition, the VFA weight ratio distribution was determined from the total VFA weight on the same day. The process stability was evaluated based on variations in daily VFA concentrations. The formula for the performance parameters is reported in Multimedia Appendix 1. The exploratory data analysis and 2-tailed *t* test on VFA data were performed for the VFA concentration, yield, and VFA/SCOD ratio for 2 HRTs by the open source program R (version 3.5.0; The R Foundation for Statistical Computing). We assumed that the 2 datasets were paired and had a normal distribution. The code is provided in the repository [27]. The values for the 2 HRTs to increase the VFA concentration in the outlet were selected based on our experience and process knowledge. According to this information, exceeding the HRT value by more than 3 - 5 days can bring the process into an anaerobic digestion step. As a result, the VFAs with high added-value markets are converted to biogas. Hence, the 2 HRTs of 3 days and 4.5 days were tried in the pilot test, knowing that the VFA concentration would either increase or decrease linearly in this local region of operation.

For the BMP tests, two kinetic models were calibrated, namely, the first-order rate and modified Gompertz, to the biogases' cumulative yield. Additionally, the specific methane production (SMP) and specific biogas production (SGP) plus maximum volumetric methane content (v/v %) were determined. Comparing these results could reveal how the biochar addition, FS/IN ratio of 0.3, and pretreatment improved the process in terms of the rate and fostered methanogenesis. Such improvements are manifested through a higher hydrolysis rate,

a shorter lag phase, and a higher maximum volumetric methane content. Besides, the biogas yield was determined as g-biogas/g-VS.

Technical and Economic Assessment

This research sets up a mass flow analysis with parameters adopted for a municipality with 70,000 population equivalents (PEs) for the two scenarios: (1) a line with pretreatment and mesophilic acidogenic fermentation followed by mesophilic anaerobic digestion and (2) a single-stage mesophilic anaerobic digestion as currently deployed at the Treviso WWTP. This study focuses on water and energy preservation and increased profits from VFA production in our conversion line through several refinements. They were tied with the HRT identified in the previous step, integration of our process knowledge of using the fine-tuned FS/IN ratio, and biochar addition in anaerobic digestion. Detailed information and calculations regarding the mass flow analysis are available in the supplementary documents in the Excel spreadsheet named “Mass Balance” [27]. The following paragraph provides the full description of the two scenarios.

The two scenarios shared the first part of the model where the separated OMSW by a door-to-door collection system that was screw-pressed and diluted with water to reach the TS of 280 g/kg. Then, in the first scenario, adding a sodium hydroxide solution (40% kg/kg) elevated the feedstock pH to 9 - 10. Afterward, the solution was heated at 60 ˚C for 24 hours in the pretreatment unit. Next, it was diluted and heated further before feeding into the mesophilic acidogenic fermenter based on the desirable HRT. The last part of the first scenario was the optimized anaerobic digestion of residual fermented solids. Specifically, the stability endowment by adding biochar to the anaerobic digestion could ultimately smooth running the process in a high OLR (low water dilution). Furthermore, an FS/IN ratio of 0.3 was applied to increase the kinetics rate with the benefit of a decrease in digester volume, energy consumption, and capital cost. This finding is of significant importance in plants and zones with limited area, water, and energy.

In the second scenario, the screw-pressed feedstock was diluted and immediately fed into a mesophilic anaerobic digester for only biogas production.

It was assumed that the reactors transfer heat from the walls with the atmosphere and earth. Further, the biogas would be consumed in the combined heat and power units for electricity production with an overall efficiency of 0.4. In this research, the mass of VFAs and the net amount of energy production were accounted for as the source of income. Meanwhile, the corresponding costs were the operational expenditure, the mass of the water process, and the final residual solids to dispose of. Reference parameters for the energy analysis and boundary conditions are given in Table 2. The price of electricity was assumed to be €130 per megawatt-hour (MWh). These two scenarios were compared to identify the most favorable one regarding surplus thermal energy and electricity or the shorter payback period.

Table . Reference parameters and boundary conditions for energy flow analysis.

Parameter	Heat transfer coefficient (W/(m^2.˚C))	Temperature (˚C)	Low heat value (MJ[a]/Nm^3)	Energy conversion efficiency
Biogas	—[b]	—	23.012	—
Thermal energy yield	—	—	—	0.5
Electrical energy yield	—	—	—	0.4
Operative temperature	—	37	—	—
Water temperature	—	15	—	—
Air temperature	—	20	—	—
Ground temperature	—	25	—	—
Outer concrete reactor wall	0.7	—	—	—
Inner concrete reactor wall	1.2	—	—	—
Floor	2.85	—	—	—

[a]MJ: megajoules.

[b]Not applicable.

Ethical Considerations

This research was not conducted on human or animal subjects and does not involve the collection of any new data. Therefore, it was unnecessary to obtain ethics approval.

Results

Biorefinery Process Scheme and Experimental Studies: Composition and Characteristics of the Pretreated Feedstock

The pretreated feedstock’s main physical and chemical characteristics were quite stable throughout the experiment (Table 3). The feedstock had an average TS content of 45 (SD 3.15) g/kg and VS content of 32 (SD 3.28) g/kg. These values

suggest that the biodegradable solids constituted 72% of the TS, which could support the fermentation process. The chemical composition of the solid part was 12.9 g-N/kg-TS, 4 g-P/kg-TS, and 565 g-COD/kg-TS, which was in the range of the values reported for the typical OMSW in Italy [30]. The chemical composition of the liquid was 325 mg N-NH_4^+/L, 14 mg P-PO_4^{3-}/L, and 25.8 g-SCOD/L. Further, the feedstock COD:N:P ratio was determined as 100:2.2:0.7, meaning that nutrients such as phosphor and nitrogen should not be the limiting substrates in acidogenic fermentation [31]. In this regard, the slight level of VFA concentration at the level of 3.5 g-SCOD/L was due to acidogenic fermentation, which had been happening during transportation.

Table . Main physical-chemical features of the feedstock.

Parameter	Weight ratio (g/kg)	Mass ratio (%)	Concentration (mg/L)
Total solids, mean (SD)[a]	45 (3.15)	—[b]	—
Volatile solids, mean (SD)[a]	32 (3.28)	—	—
Total Kjeldahl nitrogen[c]	12.9	—	—
Phosphorous[c]	4	—	—
Chemical oxygen demand[c]	565	—	—
Chemical oxygen demand:nitrogen:phosphorous	—	100:2.2:0.7	—
Soluble chemical oxygen demand	—	—	25,814
N-NH_4^+[d]	—	—	325
P-PO_4^{3-}[e]	—	—	14
Volatile fatty acid[c]	—	—	3500
Volatile solids/total solids, mean (SD)[a]	—	72 (5)	—

[a]Based on 3 measurements.

[b]Not applicable.

[c]Measurements done for nitrogen, phosphor, and soluble chemical oxygen demand equivalents for total Kjeldahl nitrogen, phosphorous, chemical oxygen demand, and volatile fatty acid.

[d]N-NH_4^+: ammonium.

[e]P-PO_4^{3-}: phosphate.

Statistical Analysis and Performance Indicators

Acidogenic Fermentation

Table 4 presents the main physical and chemical characteristics of the effluent and solid cake from the acidogenic fermenters. According to Figure 2, the process reached a steady condition after 14 days, which was roughly 3 times the HRT (4.5 days). Both HRTs were similarly stable in terms of VFA concentration variation because of a negligible difference between SDs: 2.82 g-SCOD/L versus 2.45 g-SCOD/L. These values are less than 10% of the total VFA, and the VFA production continued for more than 3 weeks without any considerable issues. The lack of any change in this process is attributed to the initial high pH of 9 - 10, which supported the process by keeping the pH variation in the optimal range of 6 - 7.5 [32].

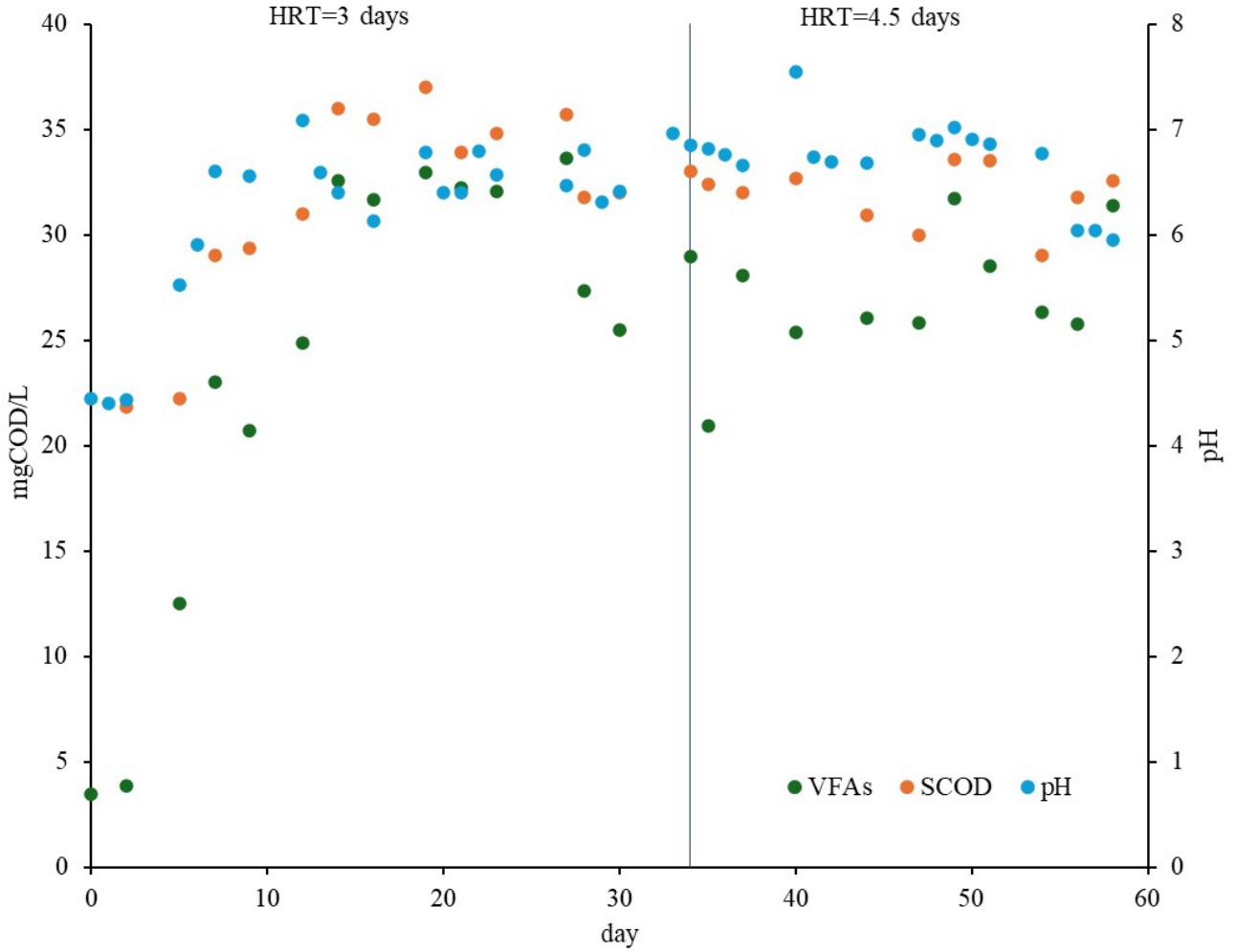
Table . Main physical-chemical features of the effluent and solid cake from mesophilic acidogenic fermentation.

Hydraulic retention time (days)	Total solids (g/kg), mean (SD)	Volatile solids (g/kg), mean (SD)	Volatile fatty acid (g–soluble chemical oxygen demand/L), mean (SD)	pH, mean (SD)
4.5[a]	43 (5.15)	23.6 (2.07)	30.77 (2.82)	6.56 (0.25)
3[b]	38 (4.55)	25.8 (1.5)	27.67 (2.45)	6.7 (0.45)

[a]5 measurements for total solids and volatile solids; 9 measurements for volatile fatty acid; 13 measurements for pH.

[b]4 measurements for total solids and volatile solids; 9 measurements for volatile fatty acid; 13 measurements for pH.

Figure 2. VFA, SCOD, and pH for mesophilic acidogenic fermentation. COD: chemical oxygen demand; HRT: hydraulic retention time; SCOD: soluble chemical oxygen demand; VFA: volatile fatty acid.



Based on the *t* test results (t_8=−2.68; *P*=.03; CI=95%), it was verified that the mean VFA concentration for an HRT of 4.5 days was significantly higher than the value for 3 days (30.77 vs 27.67 g-SCOD/L). A similar statistical analysis (t_8=−0.99; *P*=.35; CI=95%) for the VFA/SCOD ratio rejected the significance of a higher mean value of 0.892 (SD 0.04) for an HRT of 4.5 days than 3 days, with a mean value of 0.87 (SD 0.058). The possible range of values for the VFA concentrations and VFA/SCOD, which cover 99% and 50% of the data for the 2 HRTs, are depicted by the box plots in Figures 3 and 4, respectively.

Figure 3. Box plot of volatile fatty acid concentrations for mesophilic acidogenic fermentation. HRT: hydraulic retention time; SCOD: soluble chemical oxygen demand.

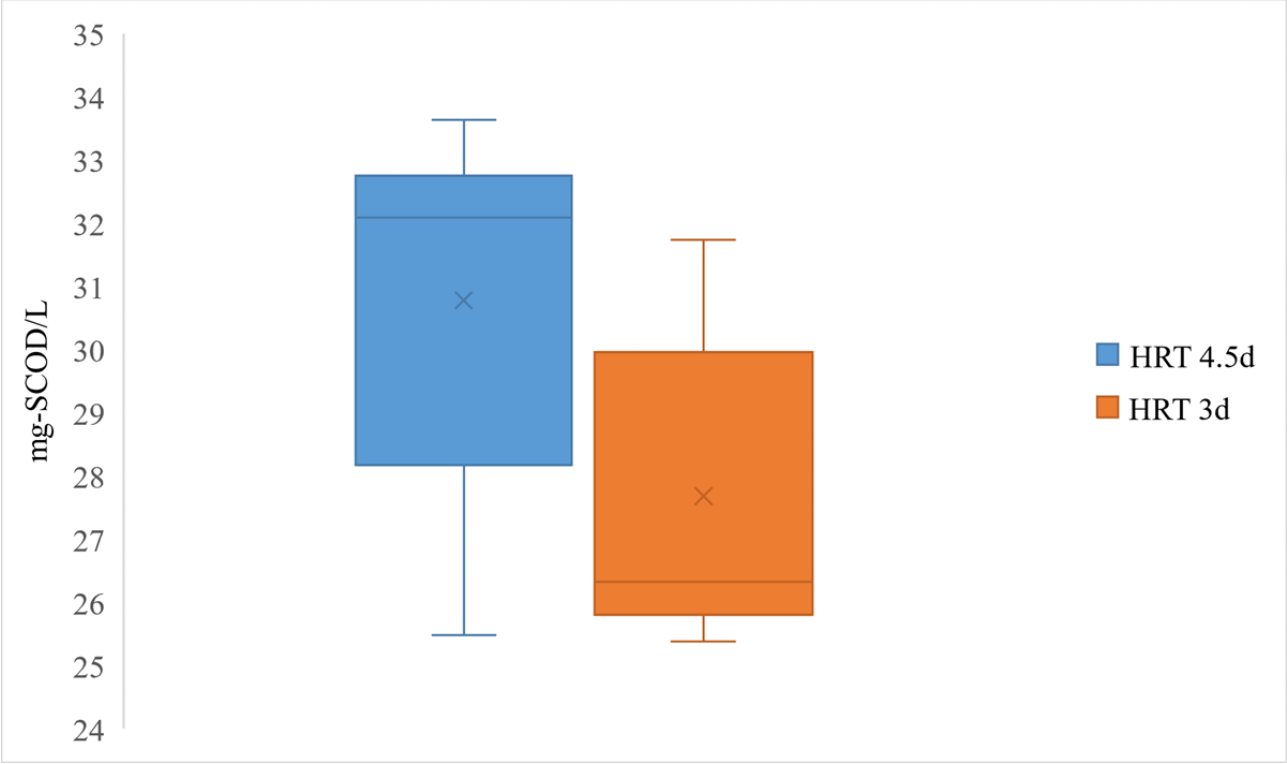
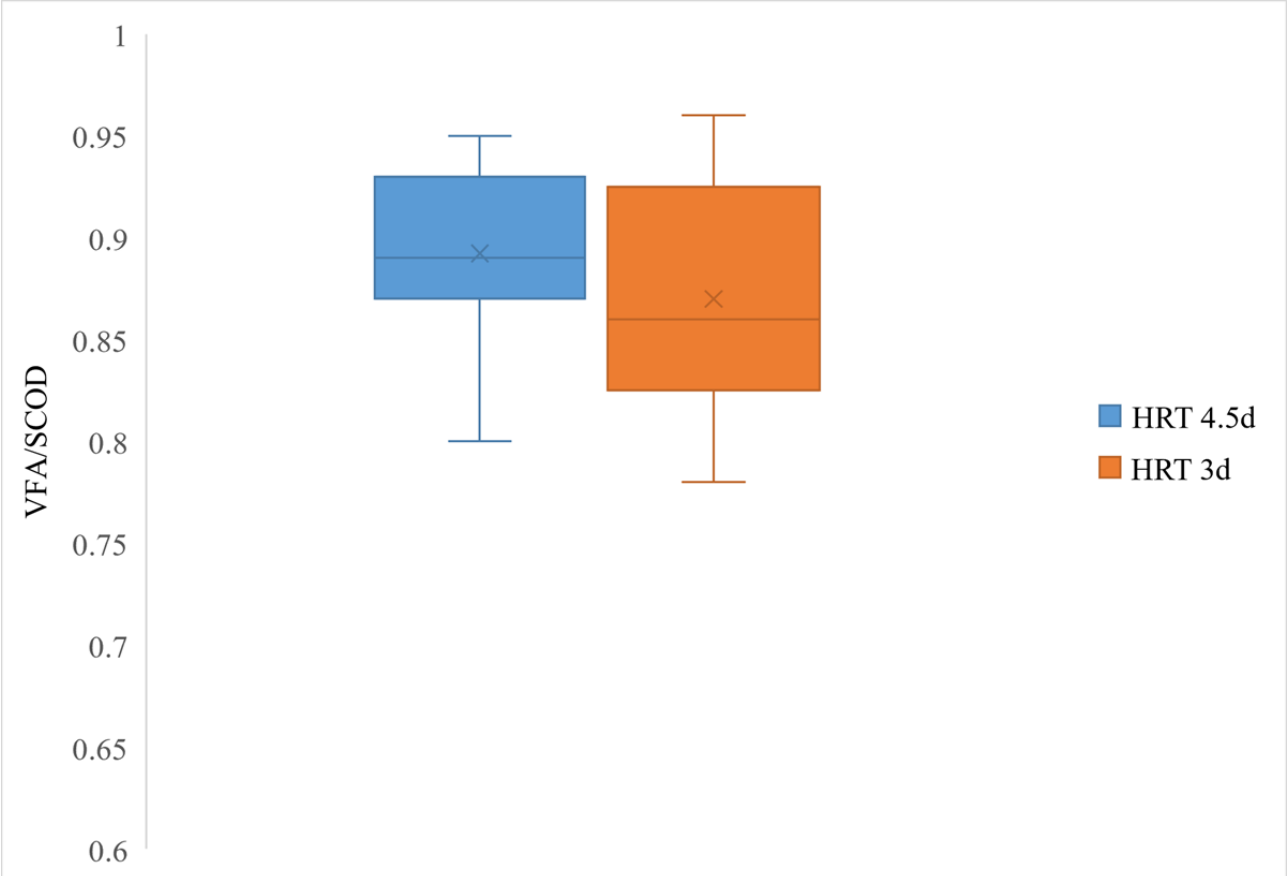


Figure 4. Box plot of VFA/SCOD ratios for mesophilic acidogenic fermentation. HRT: hydraulic retention time; SCOD: soluble chemical oxygen demand; VFA: volatile fatty acid.



Performance parameters for the 2 HRTs are given in Table 5. As can be seen, the HRT of 4.5 days gave higher COD solubilization and released more ammonia and phosphate than the HRT of 3 days. Moreover, the 0.57 VFA yield per gram of

VS for the HRT of 4.5 days was significantly higher than 0.5 for the HRT of 3 days (t_8=−2.94; P=.02; CI=95%).

In the biopolymer-synthesizing process, the aim was to generate a stable VFA weight ratio distribution with a high VFA/SCOD ratio for an efficient PHA synthesis during the whole process. Concisely, the VFA stream with a higher dominance of even numbers of carbon atom acids means a higher 3-hydroxybutyrate monomer synthesis compared to the 3-hydroxyvalerate, which is correlated with the net prevalence of odd numbers of carbon atom acids (propionic, valeric, and isovaleric acid) [33]. As can be inferred, the stability in the VFA spectrum means a predictable and reproducible PHA monomer production. Accordingly, the physical and mechanical features of synthesized biopolymers are stable [34,35].

Figure 5 reports the weight ratio distribution of the VFAs for the 2 HRTs. The main fractions were acetic acid (38% - 42%), butyric acid (24%), caproic acid (16% - 18.5%), propionic acid (9% - 11%), and valeric acid (5%). This VFA distribution, with a major part of butyric and acetic acid, is in line with those reported in similar studies [29,31]. In this respect, the VFA weight ratio distribution is determined by the type of feedstock and food waste rather than the operational conditions.

Table . Performance parameters of two different operational conditions used in mesophilic acidogenic fermentation.

Hydraulic retention time (days)	Solubilization (Δg–soluble chemical oxygen demand/g-VS^a_0), mean (SD)	Y_{VFA}[b] (Δg-VFA/g-VS_0), mean (SD)	Ammonia release (%), mean (SD)	Phosphate release (%), mean (SD)
4.5[c]	0.28 (0.06)	0.57 (0.06)	35 (10.74)	13.7 (8.77)
3[d]	0.19 (0.05)	0.50 (0.06)	29 (0.11)	11 (0.06)

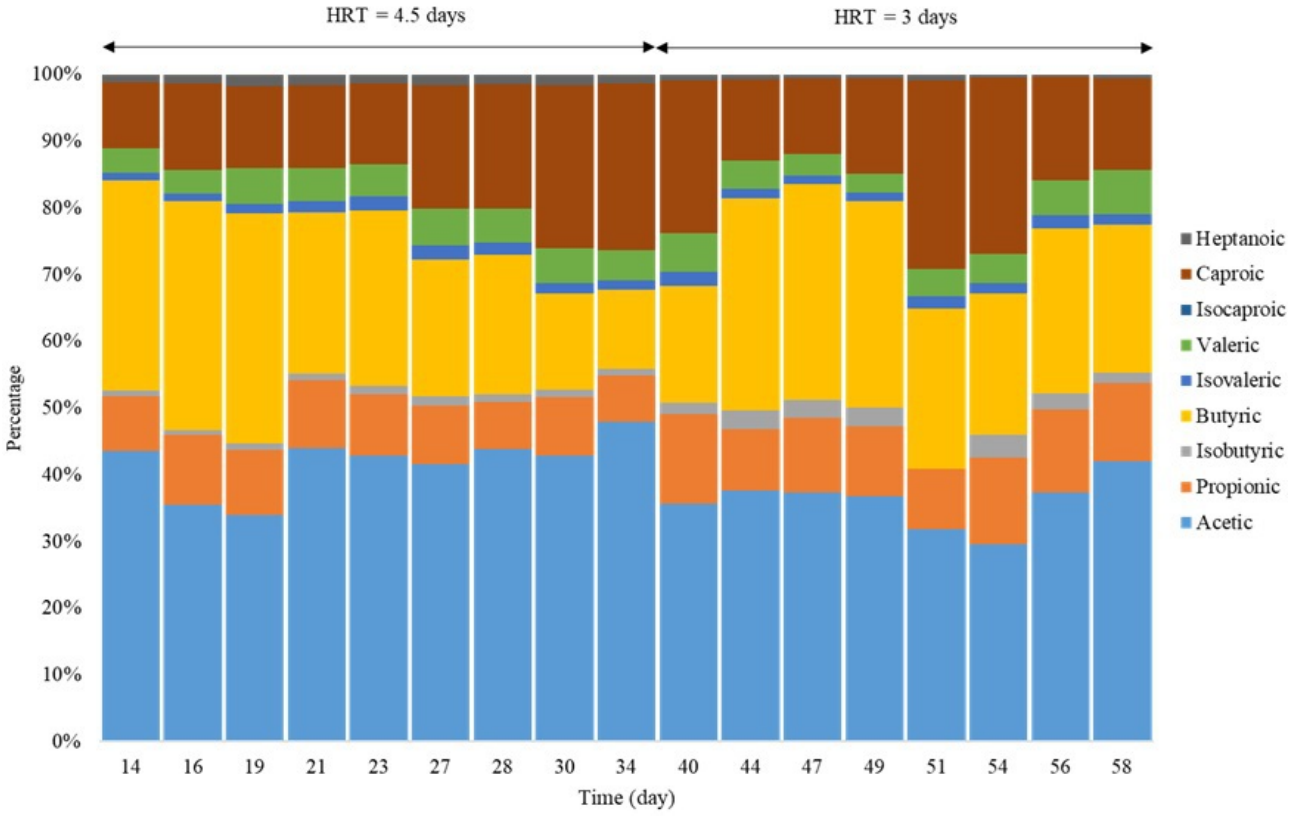
[a]VS: volatile solids.

[b]VFA: volatile fatty acid.

[c]9 measurements for solubilization (Y_{VFA}); 8 measurements for ammonia and phosphate release.

[d]9 measurements for solubilization (Y_{VFA}); 7 measurements for ammonia and phosphate release.

Figure 5. Volatile fatty acid weight ratio distribution for mesophilic acidogenic fermentation. HRT: hydraulic retention time.



Anaerobic Digestion

Table 6 summarizes the performance parameters and the results from the kinetics study for anaerobic digestion. This study obtained a remarkably high value for the hydrolysis rate (ie, 0.58, 1/d) with no lag phase. Besides, a biogas yield of 0.61 - 0.83 g-biogas/g-VS, SMP of 0.133 - 0.204 CH_4-Nm^3/kg-VS, and an average composition of 45% - 58% methane (v-CH_4/v-biogas) were obtained. According to Figure 6, adding biochar provided the desirable conditions for the growth of hydrogen using methanogenesis manifested through

a higher maximum volumetric methane content (86% vs 66% volumetric basis [v/v]).

Table . The performance indicators for anaerobic digestion and results from the kinetics study for two models: (1) first-order rate and (2) modified Gompertz.

Experiments	Specific methane production (CH_4[a]-Nm^3/kg-VS[b])	Specific gas production (CH_4-Nm^3/kg-VS)	K[c] (1/d)	R_m[d] (CH_4-mL/g-VS.d)	[e] (days)	RMSE[f] first-order (CH_4-Nm^3/kg-VS)	RMSE modified Gompertz (CH_4-Nm^3/kg-VS)	Max CH_4 content (v/v[g]), %
Without biochar	0.204	0.540	0.57	76.12	0	10.4	6.82	68.5
Biochar (0.12 g-biochar/g-)	0.133	0.567	0.69	62.42	0	5.74	5.59	86
Biochar (0.24 g-biochar/g-)	0.177	0.500	0.58	65.17	0	9.64	3.39	76.5

[a]CH_4: methane.

[b]VS: volatile solids.

[c]Hydrolysis rate.

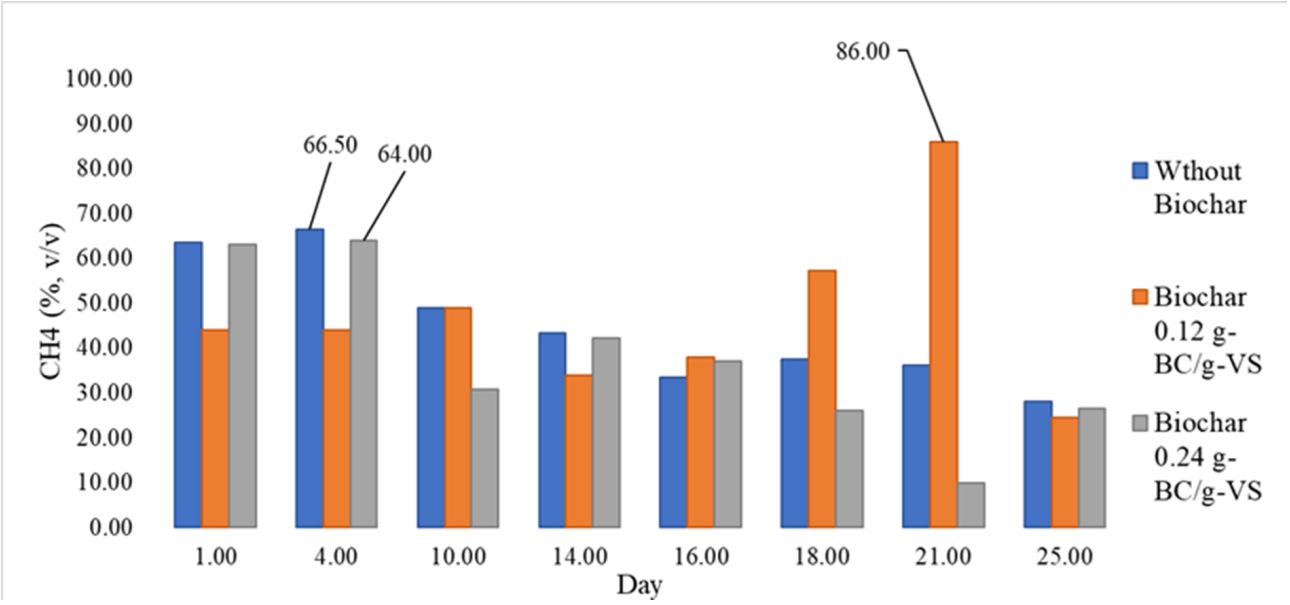
[d]Maximum methane production rate.

[e]Lag phase.

[f]RMSE: root mean squared error.

[g]v/v: volumetric basis.

Figure 6. CH_4 content in v/v for 3 different biochar dosages in anaerobic digestion. BC: biochar; CH_4: methane; VS: volatile solids; v/v: volumetric basis.



The mass flow model was adopted for 0.12 g-biochar/g-VS as the only feasible solution. Unlike other dosages, it could satisfy the assumptions for an FS/IN ratio of 0.3 at an HRT of 20 days, which was adequately long enough to let the methanogens reproduce themselves. Detailed information is available in the Excel sheet named "DIGESTER DESIGN" [27]. Besides, the high alkalinity of the biochar as reported in Table S1 in Multimedia Appendix 1 signifies a benefit of the biochar addition in limiting the concern about decreases in pH for a high OLR in full-scale implementation. Accordingly, almost 4-fold of the ordinary OLR was obtained, that is, 6.25 kg-VS/m^3.d, by minimum water dilution, knowing that the biochar could maintain the stability of the process. Therefore, the digester volume will decline at the rate of 28 L/PE. Hence, the presented mass flow line model was implemented based on the results of 0.12 g-biochar/g-VS, the weighted average composition of biomethane as 35% v/v, and the SGP as 0.56 biogas-Nm^3/kg-VS for an HRT of 20 days corresponding to an FS/IN ratio of 0.3.

Based on the root mean squared error reported in Table 6, both models were almost identical in describing biomethane production for a biochar dosage of 0.12 g-biochar/g-VS, and for simplicity, we used the first-order rate model in the feasibility study.

Technical and Economic Assessment

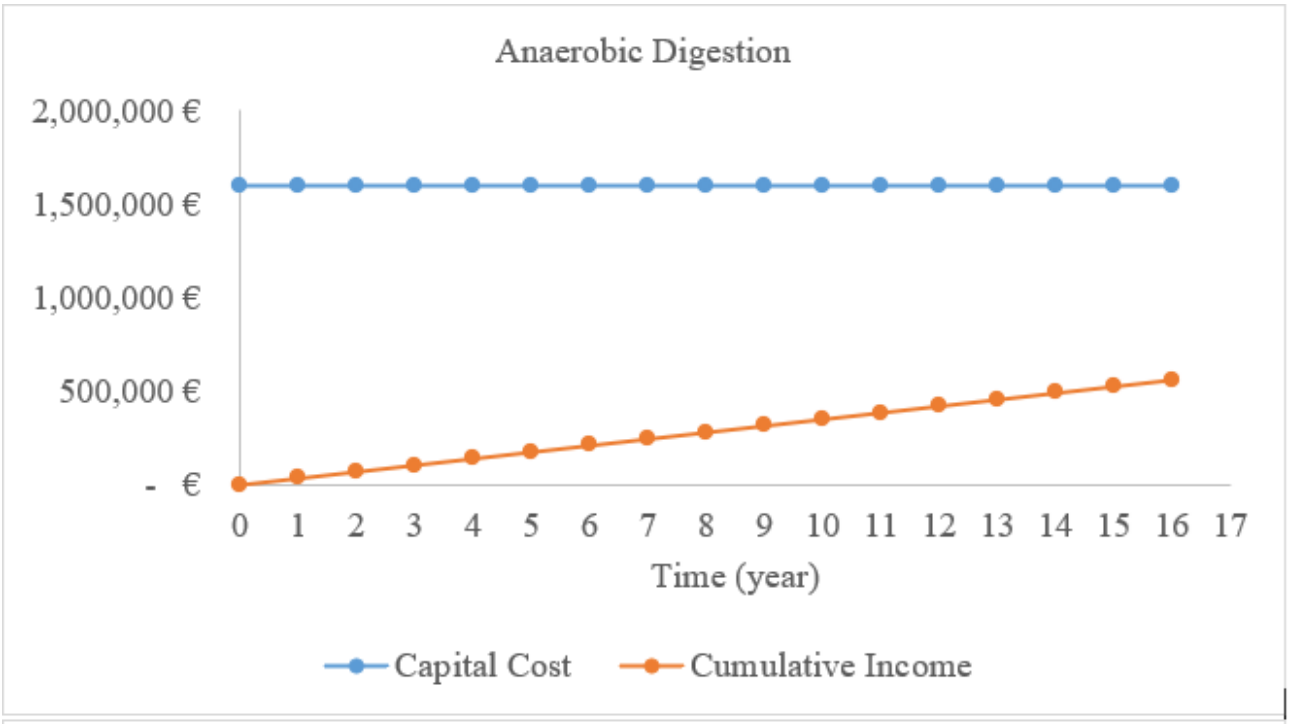
Assuming an imaginary municipality of 70,000 PEs and the amount of TS production per capita as 0.3 kg/PE per day [36], the inlet to the scale-up line would be 21,000 kg-TS per day.

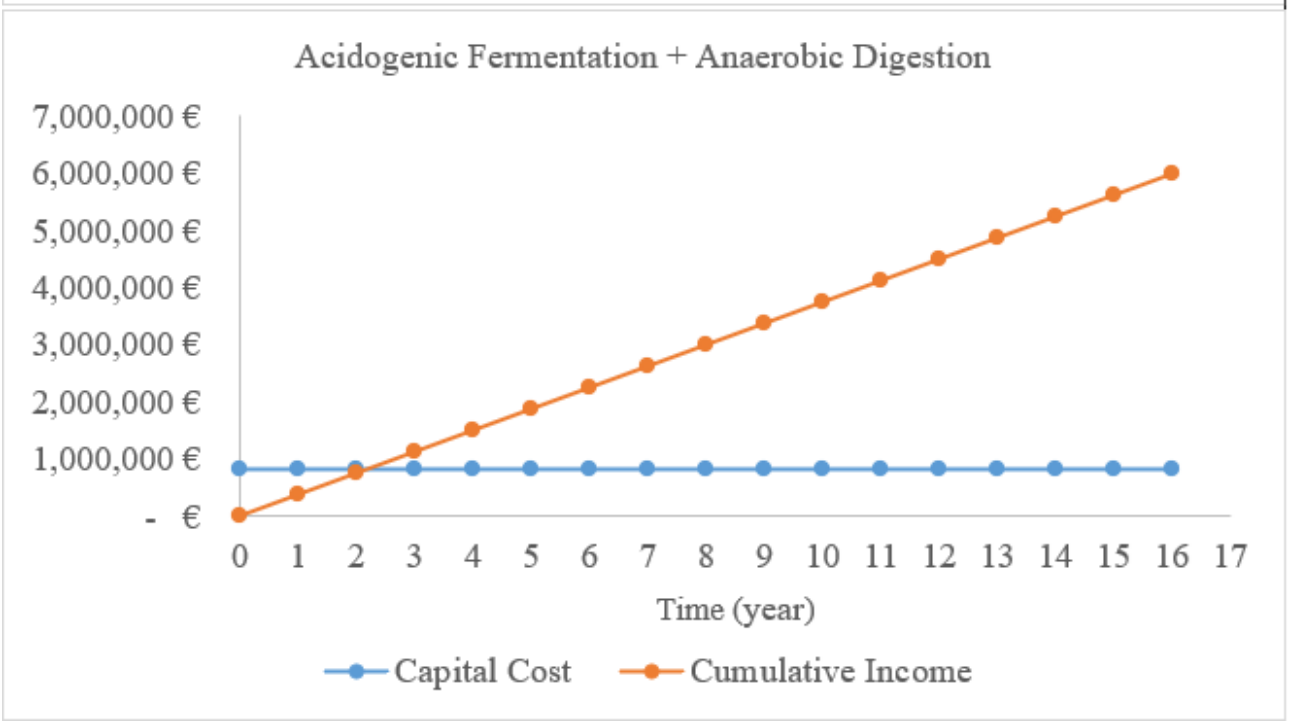
In the first scenario, the biowaste stream, after passing through the screw press and pretreatment unit, had a mass flow of 113,788 kg per day, TS of 4.1% kg/kg, and VS of 3.1% kg/kg. Then, the mixture was heated to 37 ˚C before and in the acidogenic fermenter, which was operated at an HRT of 4.5 days and OLR of 6.89 kg-VS/m^3.d. This process was performed to convert biosolids into the VFAs and SCOD at concentration levels of 30.77 g-SCOD/L and 34 g-SCOD/L, respectively. At this step, the gaseous flow rate was assumed to be zero, as an HRT of 4.5 days is short for any adequate growth of methanogens in mesophilic conditions. The stream out of the acidogenic fermenter had a mass flow rate of 113,788 kg per day, with a VFA content of 3501 kg-SCOD per day, which could be used in the PHA-synthesizing step [37]. The outlet of this step was used in the separator to gain overflow and solid cake. Later, the solid cake was minimally diluted by water before being fed into a mesophilic anaerobic digester with a biochar addition of 0.12 g-biochar/g-VS. The anaerobic digester received a TS content of 18% kg/kg and a flow rate of 18,180 kg per day, corresponding to an HRT of 20 days and OLR of 6.25 kg-VS/m^3.d. Overall, an SGP of 0.285 (Nm^3-biogas/kg-VS) was obtained assuming zero gas production in acidogenic fermentation.

In the second scenario, the fresh feedstock, after being screw-pressed, had a mass flow rate of 4678 kg-TS per day and 28% kg/kg dry matter. Then, it was diluted with water and heated before being fed into the anaerobic digester. At this step, the mass flow rate of 85,012 kg per day with a TS of 6% kg/kg entered the digester with a volume of 2125 m^3, leading to an HRT of 25 days and OLR of 1.7 kg-VS/m^3.d. The SMP of 0.311 Nm^3-biogas/kg-VS was obtained by destroying 80% of the VS.

In this study, working volumes of 512 m^3 and 364 m^3 were adopted for the acidogenic fermenter and anaerobic digester in the first scenario, respectively, and 2125 m^3 for the anaerobic digester in the second scenario. As a result, the capital cost for the presented line was almost €809,000, roughly half of the quantity for the single-step anaerobic digestion (Figure 7). Unlike the single-step anaerobic digestion that converts all VS to biogas, this novel line shared the recovery of VS between higher added-value VFAs and biogas production, and expectedly generated 10-fold higher benefits (€375,085). Consequently, the payback period was reduced by more than 20 times in 2 years (Figure 7). This period was achieved using less surplus energy (2251 megajoules [MJ]/d) for the 2-step fermentation (vs 21,567 MJ/d for the single-step anaerobic digestion).

Figure 7. Capital cost and cumulative yearly income for the two proposed scenarios. A currency exchange rate of €1=US $1.05 is applicable.





Discussion

Principal Results

We showed that multistep fermentation followed by anaerobic digestion is both economically and technically feasible. The findings indicated that producing VFAs and biogas in separate stages can significantly reduce the payback period for upcoming investments in biorefinery projects and result in the creation of a highly desired stream that is rich in VFAs. Additionally, the process stability could be maintained even at a high OLR by adding biochar and converting the VS's easily biodegradable COD content into VFAs in the first phase. This would preserve energy and water, and reduce the digester's volume.

Comparison With Previous Studies

Because of the extra pretreatment unit in this research, the VFA yield of 0.57 - 0.63 Δg-VFA/g-COD_{IN} was roughly double the value reported by Valentino et al [31] for the same OMSW.

Our results also indicate a substantial improvement in the process kinetics, which was manifested through a more than 8-fold rise in the hydrolysis rate (0.58 vs 0.07, 1/d) and a full decrease in the lag phase (0 vs 2.69 days) as opposed to the previous study by Karki et al [38]. This improvement is

attributed to the destruction of the solids structure caused by bacterial enzymes and a hot alkaline solution. Additionally, a higher active biomass per feedstock was provided using a fine-tuned FS/IN ratio of 0.3 (VS basis), which was noticeably lower than the quantities (1 and 0.5) reported in similar studies [38,39].

The values for SMP and mean methane volumetric content presented in this study are lower than those reported by Valentino et al [29] (ie, 0.25 CH_4-Nm^3/kg-VS and 63% - 64% v/v, respectively). This difference is explained by the added fresh WS, which has a higher digestible content and better nutrient balance than the fermented solids. Similarly, the SMP in this study was lower than the 0.384 CH_4-Nm^3/kg-VS found in the study by Moreno et al [39]. This study investigated the anaerobic digestion of residual solids from two steps of bioethanol production and saccharification on OMSW. In this respect, bioethanol production can only convert part of the cellulose, starch, and some dissolved carbohydrates. Consequently, a great part of the biosolids' volatile content, nearly 70%, is still available to be harvested in different biorefinery schemes compared with the one proposed in this method with 55%. Besides, the fermented OMSW would have a completely incompatible composition since it did not only come from different geographical locations (Spain and the United Kingdom) with different food habits but also underwent different biological pretreatment. Further, the multistep recovery line proposed in our study is more practicable technically. As the method studied by Moreno et al [39] requires sterilization conditions, imposing an additional operational cost and bioethanol concentration should be high enough to lower the cost of the subsequent distillation step.

Furthermore, our method for VFA production distinctively from biogas was preferable to the study by Papa et al [9], wherein the operational alteration on a single anaerobic digester was performed to obtain VFAs and biogas. These researchers asserted that the single-step recovery of biogas and VFAs was feasible by increasing the OLR while keeping the SMP of the reactor almost unaffected. The main recovery path for the VS was still biogas production in their study, which accounted for more than 90% of the VS conversion. Meanwhile, our study obtained 36% and 64% of the biogas and VFA conversion share, respectively. Further, whereas the destruction of VS of around 70% was achieved in both studies, their proposal limited the VFA distribution to propionic and butyric acid. The explanation is that some of the VFAs were converted into biogas in the same unit, which could negatively affect the PHA synthesis step later.

Conclusion and Limitations

This paper demonstrated the technical and economic feasibility of a multistep recovery line for OMSW. The results of this study indicate that the production of VFAs and biogas in distinct steps can considerably shorten the payback period for future investments in biorefinery projects and produce a highly desirable VFA-rich stream. Further, adding biochar and converting easily degradable COD content in the VS into VFAs in the first step could maintain the process stability even with a high OLR in anaerobic digestion. As a result, it leads to energy and water preservation and a decrease in the digester volume. However, consideration should be paid to the full-scale implementation since the pilot studies cannot resemble the stability of the real process. For instance, operational alterations such as raising the OLR and the addition of biochar in the full-scale implementation might perturb the process pH or the synergetic balance between the bacterial communities and stop the process completely, which was never observed in our experimental study. Further, the superb profitability of the proposed line was highly variable because our cost analysis was too simplistic and did not elaborate on all the possible associated expenditures and incomes. Besides, since many of its components were from subject matter experts rather than the pilot studies' budget, they were prone to site variations and uncertainties. Addressing the systematic uncertainty in the labor and material costs due to the changes in the supply chain issues, inflation, and site variations is beyond our scope. Moreover, caution should also be considered regarding the significance of the BMP results with the marginal difference since the number of samples was not large enough for statistical analysis. Nevertheless, the results presented in this study were prepared cautiously both technically and financially to encourage the revolution in the current state of organic waste valorization in Italy and any similar location.

In conclusion, a robust framework was proposed to assess the valorization of organic waste through experimental tests, statistical analysis, process kinetics, and mass and energy flow analysis. The findings support considerably higher profitability and, thus, a shorter payback period for the multistep fermentation than the current single anaerobic digestion. Additionally, our results encourage the circular economy perspective on converting OMSW into biogas and VFAs with the benefit of fewer residual solids due to reusing them in a pyrolysis line.

Acknowledgments

The author gratefully acknowledges the Italian Ministry of University and Research and the University of Ca′ Foscari for financially supporting and providing a study design for this research in the frame of Programma Operativo Nazionale Ricerca e Innovazione 2014-2020. In addition, the author appreciates Marco Gottardo for his helpful assistance with gas chromatography, and Francesco Valentino for his helpful comments and guidance during the study, as well as Alessio Dell ′Olivo and Aditi Parmar Chitharanjan for their helpful advice with laboratory experiments and material flow analysis. The author also attests that they used generative artificial intelligence in checking the text regarding grammar and correct word usage.

Data Availability

The codes for statistical analysis as well as the datasets generated and analyzed during this study are available from a repository [27]. Research Data Policy Type 2 (for life sciences) by Springer Nature is distributed under the terms of the Creative Commons Attribution 4.0 International License.

Conflicts of Interest

The author declares his current expert witness position as a peer reviewer in the *Journal of Medical Internet Research*.

Multimedia Appendix 1
Supplementary table, equation, digester design, and code.
[DOCX File, 31 KB - xmed_v6i1e50458_app1.docx]

References

1. Preparatory study on food waste across EU 27: final report. Publications Office of the European Union. 2011. URL: https://data.europa.eu/doi/10.2779/85947 [accessed 2024-01-29]
2. Pfaltzgraff LA, De bruyn M, Cooper EC, Budarin V, Clark JH. Food waste biomass: a resource for high-value chemicals. Green Chem 2013;15(2):307-314. [doi: 10.1039/c2gc36978h]
3. Circular economy action plan. European Commission: Environment. 2022. URL: https://ec.europa.eu/environment/strategy/circular-economy-action-plan_en#:~:text [accessed 2025-01-23]
4. Mazur-Wierzbicka E. Towards circular economy—a comparative analysis of the countries of the European Union. Resources 2021;10(5):49. [doi: 10.3390/resources10050049]
5. Mattioli A, Gatti GB, Mattuzzi GP, Cecchi F, Bolzonella D. Co-digestion of the organic fraction of municipal solid waste and sludge improves the energy balance of wastewater treatment plants: Rovereto case study. Renewable Energy 2017 Dec;113:980-988. [doi: 10.1016/j.renene.2017.06.079]
6. Cabbai V, De Bortoli N, Goi D. Pilot plant experience on anaerobic codigestion of source selected OFMSW and sewage sludge. Waste Manag 2016 Mar;49:47-54. [doi: 10.1016/j.wasman.2015.12.014] [Medline: 26739455]
7. Girotto F, Alibardi L, Cossu R. Food waste generation and industrial uses: a review. Waste Manag 2015 Nov;45:32-41. [doi: 10.1016/j.wasman.2015.06.008] [Medline: 26130171]
8. Tamisa J, Lužkov K, Jiang Y, van Loosdrecht MCM, Kleerebezem R. Enrichment of Plasticicumulans acidivorans at pilot-scale for PHA production on industrial wastewater. J Biotechnol 2014 Dec 20;192 Pt A:161-169. [doi: 10.1016/j.jbiotec.2014.10.022] [Medline: 25456060]
9. Papa G, Pepè Sciarria T, Carrara A, Scaglia B, D'Imporzano G, Adani F. Implementing polyhydroxyalkanoates production to anaerobic digestion of organic fraction of municipal solid waste to diversify products and increase total energy recovery. Bioresour Technol 2020 Dec;318:124270. [doi: 10.1016/j.biortech.2020.124270] [Medline: 33099102]
10. Jung S, Shetti NP, Reddy KR, et al. Synthesis of different biofuels from livestock waste materials and their potential as sustainable feedstocks – a review. Energy Conversion Manage 2021 May;236:114038. [doi: 10.1016/j.enconman.2021.114038]
11. Pagliano G, Ventorino V, Panico A, Pepe O. Integrated systems for biopolymers and bioenergy production from organic waste and by-products: a review of microbial processes. Biotechnol Biofuels 2017 May 2;10:113. [doi: 10.1186/s13068-017-0802-4] [Medline: 28469708]
12. Sun J, Zhang L, Loh KC. Review and perspectives of enhanced volatile fatty acids production from acidogenic fermentation of lignocellulosic biomass wastes. Bioresour Bioprocess 2021 Aug 2;8(1):68. [doi: 10.1186/s40643-021-00420-3] [Medline: 38650255]
13. Sampath P, Brijesh, Reddy KR, et al. Biohydrogen production from organic waste – a review. Chem Eng Technol 2020 Jul;43(7):1240-1248. [doi: 10.1002/ceat.201900400]
14. Gottardo M, Dosta J, Cavinato C, et al. Boosting butyrate and hydrogen production in acidogenic fermentation of food waste and sewage sludge mixture: a pilot scale demonstration. J Cleaner Production 2023 Jun 10:136919. [doi: 10.1016/j.jclepro.2023.136919]
15. Micolucci F, Gottardo M, Bolzonella D, Pavan P, Majone M, Valentino F. Pilot-scale multi-purposes approach for volatile fatty acid production, hydrogen and methane from an automatic controlled anaerobic process. J Cleaner Production 2020 Dec 20;277:124297. [doi: 10.1016/j.jclepro.2020.124297]
16. Anderottola G, Canziani R, Foladori P, Ragazzi M, Tatano F. Laboratory scale experimentation for RBCOD production from OFMSW for BNR systems: results and kinetics. Environ Technol 2000 Dec;21(12):1413-1419. [doi: 10.1080/09593332208618173]
17. Stoyanova E, Lundaa T, Bochmann G, Fuchs W. Overcoming the bottlenecks of anaerobic digestion of olive mill solid waste by two-stage fermentation. Environ Technol 2017 Feb;38(4):394-405. [doi: 10.1080/09593330.2016.1196736] [Medline: 27279450]

18. Sauer M, Porro D, Mattanovich D, Branduardi P. Microbial production of organic acids: expanding the markets. Trends Biotechnol 2008 Feb;26(2):100-108. [doi: 10.1016/j.tibtech.2007.11.006] [Medline: 18191255]
19. Dahiya S, Sarkar O, Swamy YV, Venkata Mohan S. Acidogenic fermentation of food waste for volatile fatty acid production with co-generation of biohydrogen. Bioresour Technol 2015 Apr;182:103-113. [doi: 10.1016/j.biortech.2015.01.007] [Medline: 25682230]
20. Ma Y, Gu J, Liu Y. Evaluation of anaerobic digestion of food waste and waste activated sludge: soluble COD versus its chemical composition. Sci Total Environ 2018 Dec 1;643:21-27. [doi: 10.1016/j.scitotenv.2018.06.187] [Medline: 29935360]
21. Montalvo S, Vielma S, Borja R, Huiliñir C, Guerrero L. Increase in biogas production in anaerobic sludge digestion by combining aerobic hydrolysis and addition of metallic wastes. Renewable Energy 2018 Aug;123:541-548. [doi: 10.1016/j.renene.2018.02.004]
22. Kumar M, Xiong X, Sun Y, et al. Critical review on biochar‐supported catalysts for pollutant degradation and sustainable biorefinery. Adv Sustainable Syst 2020 Oct;4(10):1900149. [doi: 10.1002/adsu.201900149]
23. Kumar M, Dutta S, You S, et al. A critical review on biochar for enhancing biogas production from anaerobic digestion of food waste and sludge. J Clean Prod 2021 Oct 10;305:127143. [doi: 10.1016/j.jclepro.2021.127143] [Medline: 36570877]
24. Inyang M, Gao B, Pullammanappallil P, Ding W, Zimmerman AR. Biochar from anaerobically digested sugarcane bagasse. Bioresour Technol 2010 Nov;101(22):8868-8872. [doi: 10.1016/j.biortech.2010.06.088] [Medline: 20634061]
25. Lipps WC, Braun-Howland EB, Baxter TE, editors. Standard Methods for the Examination of Water and Wastewater, 24th edition: American Public Health Association; 2022.
26. Cinar S, Cinar S, Kuchta K. Machine learning algorithms for temperature management in the anaerobic digestion process. Fermentation 2022 Jan 30;8(2):65. [doi: 10.3390/fermentation8020065]
27. Borhany H. Supplementary documents for converting organic municipal solid waste into volatile fatty acids and biogas: experimental pilot and batch studies with statistical analysis. Zenodo. 2024 Jul 14. URL: https://zenodo.org/records/12739504 [accessed 2025-01-23]
28. Mumme J, Srocke F, Heeg K, Werner M. Use of biochars in anaerobic digestion. Bioresour Technol 2014 Jul;164:189-197. [doi: 10.1016/j.biortech.2014.05.008] [Medline: 24859210]
29. Valentino F, Moretto G, Gottardo M, Pavan P, Bolzonella D, Majone M. Novel routes for urban bio-waste management: a combined acidic fermentation and anaerobic digestion process for platform chemicals and biogas production. J Cleaner Production 2019 May 20;220:368-375. [doi: 10.1016/j.jclepro.2019.02.102]
30. Moretto G, Valentino F, Pavan P, Majone M, Bolzonella D. Optimization of urban waste fermentation for volatile fatty acids production. Waste Manag 2019 Jun 1;92:21-29. [doi: 10.1016/j.wasman.2019.05.010] [Medline: 31160023]
31. Valentino F, Gottardo M, Micolucci F, et al. Organic fraction of municipal solid waste recovery by conversion into added-value polyhydroxyalkanoates and biogas. ACS Sustainable Chem Eng 2018 Dec 3;6(12):16375-16385. [doi: 10.1021/acssuschemeng.8b03454]
32. Gottardo M, Micolucci F, Bolzonella D, Uellendahl H, Pavan P. Pilot scale fermentation coupled with anaerobic digestion of food waste - effect of dynamic digestate recirculation. Renewable Energy 2017 Dec;114:455-463. [doi: 10.1016/j.renene.2017.07.047]
33. Estévez-Alonso Á, Pei R, van Loosdrecht MCM, Kleerebezem R, Werker A. Scaling-up microbial community-based polyhydroxyalkanoate production: status and challenges. Bioresour Technol 2021 May;327:124790. [doi: 10.1016/j.biortech.2021.124790] [Medline: 33582521]
34. Morgan-Sagastume F, Hjort M, Cirne D, et al. Integrated production of polyhydroxyalkanoates (PHAs) with municipal wastewater and sludge treatment at pilot scale. Bioresour Technol 2015 Apr;181:78-89. [doi: 10.1016/j.biortech.2015.01.046] [Medline: 25638407]
35. Valentino F, Morgan-Sagastume F, Campanari S, Villano M, Werker A, Majone M. Carbon recovery from wastewater through bioconversion into biodegradable polymers. N Biotechnol 2017 Jul 25;37(Pt A):9-23. [doi: 10.1016/j.nbt.2016.05.007] [Medline: 27288751]
36. Metcalf & Eddy, Inc, Tchobanoglous G, Stensel H, Tsuchihashi R, Burton F. Wastewater Engineering: Treatment, Disposal, and Reuse, 5th edition: McGraw-Hill Education; 2014.
37. Tamis J, Joosse BM, Loosdrecht MCMV, Kleerebezem R. High-rate volatile fatty acid (VFA) production by a granular sludge process at low pH. Biotechnol Bioeng 2015 Nov;112(11):2248-2255. [doi: 10.1002/bit.25640] [Medline: 25950759]
38. Karki R, Chuenchart W, Surendra KC, Sung S, Raskin L, Khanal SK. Anaerobic co-digestion of various organic wastes: kinetic modeling and synergistic impact evaluation. Bioresour Technol 2022 Jan;343:126063. [doi: 10.1016/j.biortech.2021.126063] [Medline: 34619321]
39. Moreno AD, Magdalena JA, Oliva JM, et al. Sequential bioethanol and methane production from municipal solid waste: an integrated biorefinery strategy towards cost-effectiveness. Process Safety Environ Protection 2021 Feb;146:424-431. [doi: 10.1016/j.psep.2020.09.022]

Abbreviations

BMP: bio-methane potential

CH_4: methane
CO_2: carbon dioxide
COD: chemical oxygen demand
FS/IN: feedstock/inoculum
H_2: hydrogen molecule
HRT: hydraulic retention time
kPa: kilopascal
MJ: megajoules
MWh: megawatt-hour
N-NH_4^+: ammonium
N_2: nitrogen molecule
NaOH: sodium hydroxide
O_2: oxygen molecule
OLR: organic loading rate
OMSW: organic municipal solid waste
P: phosphorous
P-PO_4^{3-}: phosphate
PE: population equivalent
PHA: polyhydroxyalkanoate
rpm: rounds per minute
SCOD: soluble chemical oxygen demand
SGP: specific biogas production
SMP: specific methane production
TS: total solids
v/v: volumetric basis
v/v %: maximum volumetric methane content
VFA: volatile fatty acid
VS: volatile solids
WS: waste sludge
WWTP: wastewater treatment plant

Edited by T Leung; submitted 01.07.23; peer-reviewed by D Elsalamony, Anonymous; revised version received 08.07.24; accepted 12.07.24; published 04.02.25.

Please cite as:
Borhany H
Converting Organic Municipal Solid Waste Into Volatile Fatty Acids and Biogas: Experimental Pilot and Batch Studies With Statistical Analysis
JMIRx Med 2025;6:e50458
URL: https://xmed.jmir.org/2025/1/e50458
doi:10.2196/50458

Impact of the COVID-19 Pandemic on Routine Childhood Vaccination Coverage in Ecuador From 2019 to 2021: Comparative Analysis

Jose Sanchez[1*], MSc, MD; Alejandro Arjuna Rodriguez Sr[2*]; Kimberlly Pamela Montenegro Cuello Sr[2]

[1]Faculty of Health Sciences and Human Well-being, Universidad Indoamérica, Avenida Machala y Sabanilla, La Pradera, Quito, Ecuador

[2]Faculty of Health Sciences "Eugenio Espejo", Universidad UTE, Quito, Ecuador

[*]these authors contributed equally

Corresponding Author:
Jose Sanchez, MSc, MD
Faculty of Health Sciences and Human Well-being, Universidad Indoamérica, Avenida Machala y Sabanilla, La Pradera, Quito, Ecuador

Related Articles:

Companion article: https://www.medrxiv.org/content/10.1101/2025.03.26.25324742v1

Companion article: https://med.jmirx.org/2025/1/e84847

Companion article: https://med.jmirx.org/2025/1/e84848

Companion article: https://med.jmirx.org/2025/1/e84849

Companion article: https://med.jmirx.org/2025/1/e84851

Abstract

Background: The COVID-19 pandemic disrupted essential health care services globally, including routine childhood immunization programs. Ecuador faced significant challenges in maintaining vaccination coverage during this period.

Objective: The aim of this study is to analyze the impact of the COVID-19 pandemic on routine childhood vaccination coverage in Ecuador by comparing prepandemic (2019) and pandemic (2020 - 2021) data.

Methods: This retrospective observational study analyzed vaccination coverage data from the Ministry of Public Health of Ecuador and demographic data from the National Institute of Statistics and Censuses. We examined routine childhood vaccination coverage for children under 24 months across all 24 provinces. Statistical analyses were performed using SPSS (version 28.0), including descriptive statistics and comparative analysis. Coverage rates were calculated as percentages of children in target age groups receiving recommended doses.

Results: A significant decline in routine childhood vaccination coverage was observed during the pandemic. BCG vaccine coverage decreased from 86.4% in 2019 (n=286,569) to 80.7% in 2020 (n=266,961) and 75.3% in 2021 (n=248,812). Pentavalent vaccine third dose coverage dropped from 85.0% to 68.0% across the same period. The most dramatic decline was seen in measles-mumps-rubella vaccine second dose coverage, falling from 75.7% in 2019 to 58.4% in 2021. Coastal and highland provinces experienced the most severe reductions, with approximately 137,000 fewer doses administered in 2020 compared to stable prepandemic levels.

Conclusions: The COVID-19 pandemic significantly impacted routine childhood vaccination coverage in Ecuador, with sustained declines through 2021. Regional disparities were evident, with vulnerable populations facing greater challenges accessing immunization services. Urgent interventions, including catch-up campaigns and strengthened health systems, are needed to restore coverage levels and prevent outbreaks of vaccine-preventable diseases.

(JMIRx Med 2025;6:e75293) doi:10.2196/75293

KEYWORDS

COVID-19 pandemic; vaccination coverage; Ecuador; immunization; routine vaccination; health disparities; vaccine hesitancy

Introduction

Background and Global Context

The COVID-19 pandemic emerged as an unprecedented global health crisis, fundamentally disrupting health care systems and essential services worldwide [1]. Beyond the direct morbidity and mortality caused by SARS-CoV-2, the pandemic created far-reaching consequences for routine health care delivery, particularly impacting childhood immunization programs that are critical for preventing infectious diseases and maintaining population health [2,3].

Global evidence demonstrates substantial disruptions to vaccination services during the pandemic. The World Health Organization reported that at least 68% of countries experienced disruptions to childhood immunization programs, with low- and middle-income countries disproportionately affected [4]. These disruptions resulted from multiple factors including health care worker redeployment, supply chain interruptions, physical distancing measures, and reduced health care–seeking behavior due to fear of COVID-19 transmission [5,6].

Impact on Low- and Middle-Income Countries

In low- and middle-income countries, where health care infrastructure may be fragile and resources limited, the pandemic exacerbated preexisting challenges in vaccination delivery [7]. Countries in Latin America and the Caribbean faced particular difficulties, with studies showing that COVID-19 containment measures led to significant reductions in routine immunization coverage across the region [8]. Castro-Aguirre et al [9] conducted a comprehensive analysis of 39 countries and territories in Latin America and the Caribbean, finding significant reductions in diphtheria-pertussis-tetanus (DTP) vaccine coverage in 79% of assessed regions.

Ecuador's Prepandemic Vaccination Context

Ecuador, a South American country with diverse geographical regions and varying levels of health care access, operated a national immunization program that faced coverage challenges even before the pandemic [10]. The country's immunization system demonstrated disparities across different geographical regions and socioeconomic groups, with rural and Indigenous populations often experiencing lower vaccination rates [11].

Prior to 2020, Ecuador's routine childhood vaccination program included vaccines against tuberculosis (BCG), diphtheria-pertussis-tetanus-hepatitis B-*Haemophilus influenzae* type b (pentavalent), pneumococcal disease, poliovirus, rotavirus, measles-mumps-rubella, yellow fever, and varicella [12]. Coverage rates varied significantly across provinces, reflecting the country's geographical challenges and socioeconomic disparities [13].

Pandemic Impact in Ecuador

As of July 2021, only 57% of Ecuador's population had received the first COVID-19 vaccine dose, highlighting significant challenges in reaching underserved populations in remote areas [14]. The pandemic's impact on routine childhood vaccination was particularly concerning, given the potential for vaccine-preventable disease outbreaks in an already vulnerable population [15].

The implementation of movement restrictions, health care system overwhelm, and resource reallocation to COVID-19 response efforts created substantial barriers to routine immunization services [16]. Health care facilities experienced reduced capacity, parents delayed or avoided medical visits due to infection fears, and supply chains faced significant disruptions [17].

Study Rationale and Objectives

Understanding the specific impact of COVID-19 on Ecuador's childhood vaccination program is crucial for developing targeted interventions to restore coverage levels and prevent future disruptions [18]. This analysis provides essential data for policymakers and public health officials working to strengthen immunization systems and improve pandemic preparedness [19].

The primary objective of this study is to quantify the impact of the COVID-19 pandemic on routine childhood vaccination coverage in Ecuador by comparing coverage rates before (ie, 2019) and during (ie, 2020 - 2021) the pandemic and to identify geographical disparities in vaccination access during this period.

Methods

Study Design

This study used a retrospective, observational design to analyze vaccination coverage data from Ecuador's national immunization program. We conducted a comparative analysis examining routine childhood vaccination coverage for the prepandemic period (2019) and the pandemic period (2020 - 2021) [20]. This design allowed for the examination of temporal trends and changes in vaccination coverage, providing insights into the pandemic's impact on immunization services.

Data Sources and Collection

Primary data for this study were obtained from three key sources:

- Ministry of Public Health National Immunization Strategy Bulletin: This official source provided comprehensive vaccination coverage data at the national and provincial level, including the number of doses administered, target populations, and calculated coverage rates for all routine childhood vaccines [21].
- National Institute of Statistics and Censuses (INEC): INEC provided demographic data including population estimates, birth rates, and population projections used to calculate coverage rates and understand target populations [22].
- Published literature: To provide additional context and support for the findings, we conducted a systematic review of relevant peer-reviewed studies using PubMed, Scopus, and Web of Science databases with keywords including "COVID-19," "vaccination coverage," "Ecuador," "childhood immunization," and "pandemic impact" [23].

Study Population

The study population consisted of children under 24 months of age in Ecuador, representing the target age group for routine childhood vaccinations according to the national immunization schedule [24]. Data were analyzed for all 24 provinces across 4 geographical regions: Costa (coast), Sierra (highlands), Amazonía (Amazon region), and Insular (Galápagos Islands) [25].

Vaccination Coverage Metrics

We analyzed coverage for the following vaccines according to Ecuador's national immunization schedule [26]:

- BCG: administered at birth
- Hepatitis B: first dose at birth
- Pentavalent (DTP-hepatitis B-*Haemophilus influenzae* type b): three doses at 2, 4, and 6 months
- Pneumococcal conjugate: three doses at 2, 4, and 6 months
- Inactivated poliovirus vaccine: two doses at 2 and 4 months
- Bivalent oral polio vaccine: doses at 6 and ≥12 months
- Rotavirus: two doses at 2 and 4 months
- Measles-mumps-rubella (MMR): two doses at 12 and 18 months
- Yellow fever: single dose at 12 months
- Varicella: single dose at 15 months
- DTP booster: fourth dose at 12 - 15 months

Coverage rates were calculated as the percentage of children in the target age group who received the recommended number of doses for each vaccine, following World Health Organization (WHO) guidelines for vaccination coverage assessment [27].

Data Analysis

Statistical analyses were performed using SPSS (version 28.0; IBM Corp) [28]. The following analytical approaches were used:

- Descriptive statistics: We calculated frequencies, percentages, means, and standard deviations to summarize vaccination coverage data and population characteristics [29].
- Comparative analysis: Coverage rates were compared between the prepandemic year (2019) and pandemic years (2020 - 2021) using appropriate statistical methods. We calculated absolute and relative changes in coverage between time periods [30].
- Geographical analysis: We examined regional and provincial variations in vaccination coverage to identify areas most affected by pandemic-related disruptions [31].
- Trend visualization: Coverage data were plotted over time to visualize trends and identify patterns of decline or recovery across different vaccines and regions using the *matplotlib* and *seaborn* libraries in Python [32].

Ethical Considerations

This study used secondary, publicly available data from official government sources and did not involve direct human subjects research. Therefore, ethical approval from an institutional review board was not required [33]. All data were anonymized and analyzed in aggregate form, ensuring privacy protection [34].

Data Quality and Limitations

Data quality was ensured through cross-referencing between the Ministry of Public Health and INEC sources. Limitations include the lack of detailed socioeconomic data at the individual level and the absence of data beyond 2021, which would allow assessment of recovery efforts [35].

Results

Overall Vaccination Coverage Trends

Analysis of routine childhood vaccination data revealed a clear pattern of declining coverage between 2019 and 2021, demonstrating the significant impact of the COVID-19 pandemic on adherence to immunization schedules [36]. Table 1 presents comprehensive coverage data showing this concerning trend across all major vaccines.

Table . Regional and provincial population data for the years 2019, 2020, and 2021.[a]

Region and province		2019	2020	2021
Costa				
	Esmeraldas	13,293	13,211	13,128
	Manabí	29,299	29,207	29,005
	Los Ríos	18,897	18,888	18,798
	Santa Elena	8834	8897	8900
	Guayas	79,543	79,535	79,519
	Santo Domingo	10,535	10,537	10,541
	El Oro	12,526	12,464	12,438
Sierra				
	Azuay	15,903	15,700	15,688
	Bolívar	4338	4223	4205
	Cañar	5680	5660	5640
	Carchi	3258	3236	3214
	Cotopaxi	10,355	10,304	10,293
	Chimborazo	9853	9764	9660
	Imbabura	9173	9141	9115
	Loja	9978	9923	9872
	Pichincha	56,698	57,062	57,200
	Tungurahua	10,166	10,111	10,069
Amazonía				
	Morona Santiago	4895	4842	4822
	Napo	3341	3361	3381
	Orellana	3883	3821	3800
	Pastaza	2639	2659	2679
	Sucumbíos	4944	4958	4978
	Zamora Chinchipe	2839	2837	2833
Insular				
	Galápagos	624	631	666
Total		331,494	330,972	330,444

[a]Dirección Nacional de Estadística y Análisis de Información de Salud (DNEAIS), early and late capture base, developed by the Ministry of Public Health National Immunization Strategy.

Figure 1 illustrates the temporal trends in vaccination coverage for key vaccines from 2019 to 2021. The visualization clearly demonstrates the progressive decline in coverage rates, with the most dramatic decreases occurring between 2020 and 2021. A heatmap provides an alternative visualization of the comprehensive data presented in Table 2, highlighting the widespread nature of the coverage decline (Figure 2).

Figure 1. Vaccination coverage trends in Ecuador (2019-2021). The line graph shows the temporal trends for BCG, pentavalent 3, pneumococcal 3, rotavirus 2, and MMR 2 vaccines, with the 80% World Health Organization threshold line. MMR: measles-mumps-rubella.

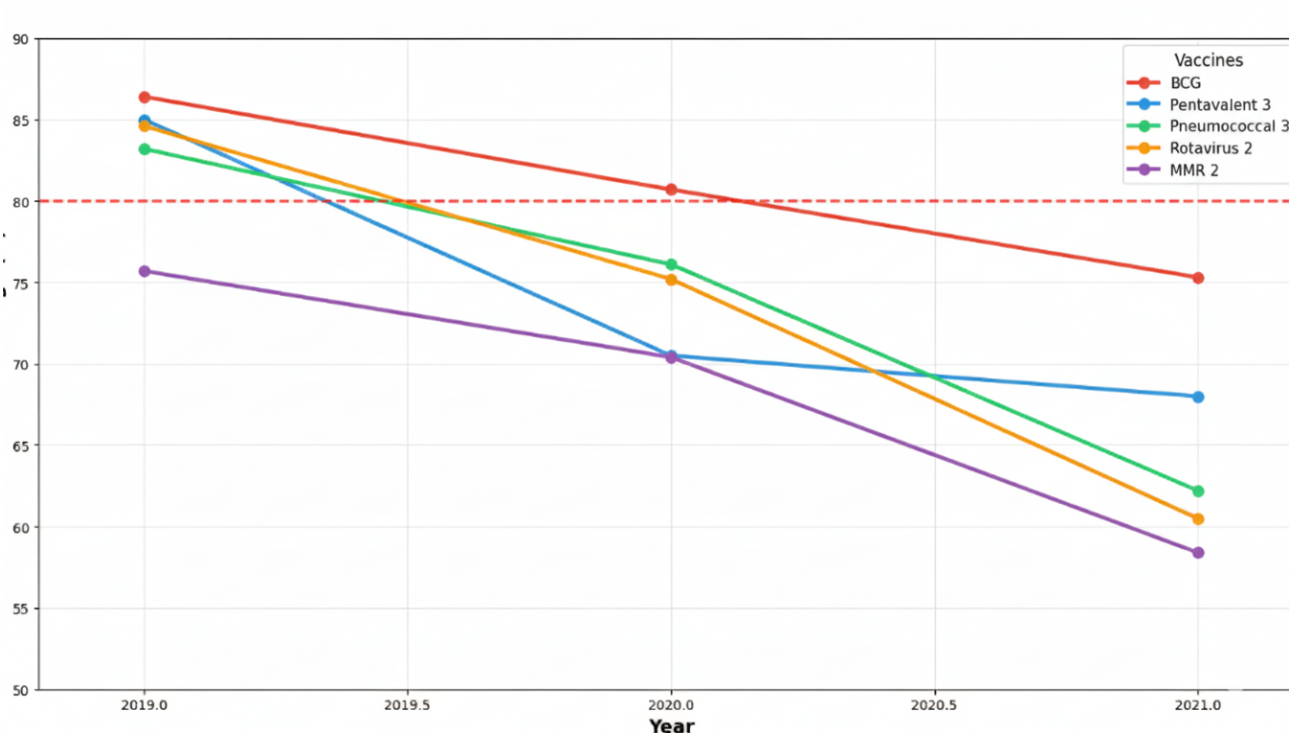


Table . Vaccination coverage by target group, vaccine type, and year (2019-2021).[a]

Target group and vaccine	2019		2020		2021	
	Doses applied	Coverage, %	Doses applied	Coverage, %	Doses applied	Coverage, %
Birth (4 h)						
BCG total	286,569	86.4	266,961	80.7	248,812	75.3
HB[b] zero	237,145	71.5	204,979	61.9	202,679	61.3
2 months						
Pentavalent 1	282,623	85.3	246,141	74.4	254,565	77
Pneumococcal 1	277,310	83.7	265,924	80.4	238,605	72.2
IPV[c] 1	282,277	85.2	263,867	79.7	232,631	70.4
Rotavirus 1	278,994	84.2	253,192	76.5	214,668	65
4 months						
Pentavalent 2	284,078	85.7	243,317	73.5	243,082	73.6
Pneumococcal 2	278,085	83.9	256,408	77.5	228,686	69.2
IPV 2	282,171	85.1	260,538	78.7	211,797	64.1
Rotavirus 2	280,431	84.6	248,973	75.2	199,909	60.5
6 months						
Pentavalent 3	281,734	85	233,371	70.5	224,702	68
Pneumococcal 3	275,947	83.2	251,977	76.1	205,659	62.2
bOPV[d] 3	280,390	84.6	239,889	72.5	193,510	58.6
12 months						
MMR 1	276,289	83.3	266,550	80.5	215,874	65.3
Yellow fever	279,008	84.2	263,123	79.5	230,524	69.8
15 months						
Varicella	268,434	81	259,880	78.5	218,800	66.2
18 months						
MMR[e] 2	250,964	75.7	232,883	70.4	192,835	58.4
1 year from third dose						
bOPV 4	254,395	76.7	229,210	69.3	193,234	58.5
DTP[f] 4	254,256	76.7	249,857	75.5	196,616	59.5

[a]Dirección Nacional de Estadística y Análisis de Información de Salud (DNEAIS), early and late capture base, developed by the Ministry of Public Health national immunization strategy.

[b]HB: hepatitis B.

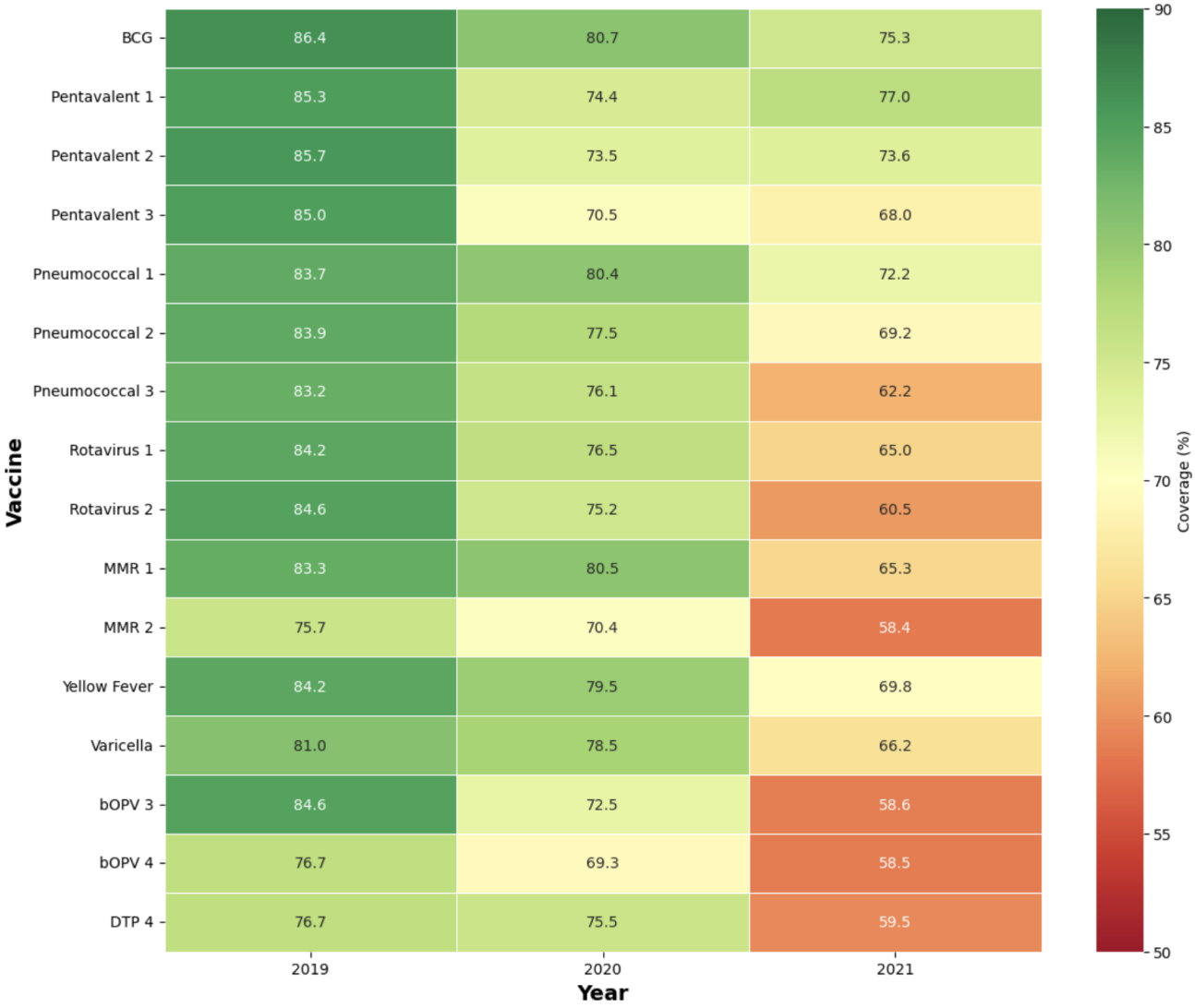
[c]IPV: inactivated poliovirus vaccine.

[d]bOPV: bivalent oral polio vaccine.

[e]MMR: measles-mumps-rubella.

[f]DTP: diphtheria-pertussis-tetanus.

Figure 2. Heatmap of vaccination coverage by vaccine and year. A color-coded heatmap showing all vaccines across the 3 years, with coverage percentages displayed. bOPV: bivalent oral polio vaccine; DTP: diphtheria-pertussis-tetanus; MMR: measles-mumps-rubella.

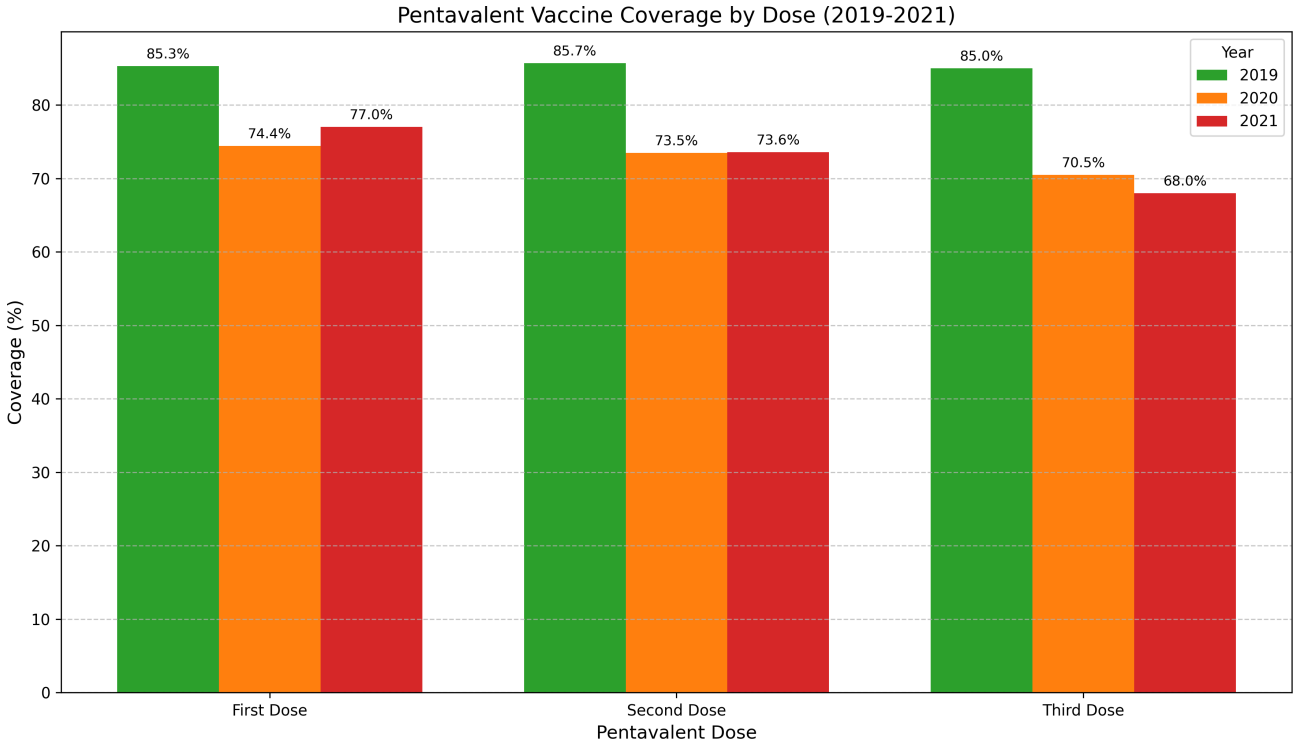


Vaccine-Specific Coverage Analysis

We conducted an analysis on vaccine-specific coverage and found the following:

- BCG vaccine (birth): Coverage for BCG, administered at birth, decreased progressively from 86.4% in 2019 (286,569 doses) to 80.7% in 2020 (266,961 doses) and further to 75.3% in 2021 (248,812 doses). This represents a cumulative decrease of 11.1 percentage points over the 2-year period [37].
- Pentavalent vaccine series: The pentavalent vaccine showed variable patterns across doses. First dose coverage declined from 85.3% in 2019 to 74.4% in 2020 but showed slight recovery to 77% in 2021. However, completion rates for the 3-dose series remained substantially below prepandemic levels, with third dose coverage falling from 85% in 2019 to 68% in 2021 [38]. This growing gap between initiation and completion of the series is a critical indicator of service disruption (Figure 3).
- Pneumococcal vaccine: This vaccine experienced consistent declines across all 3 doses. First dose coverage fell from 83.7% in 2019 to 72.2% in 2021, while third dose coverage dropped more dramatically from 83.2% to 62.2% over the same period [39].
- Rotavirus vaccine: Among the most affected vaccines, rotavirus coverage showed severe declines. Second dose coverage plummeted from 84.6% in 2019 to 60.5% in 2021, representing a 24.1 percentage point decrease [40].
- MMR: MMR vaccine coverage demonstrated significant drops, particularly for the second dose administered at 18 months. Coverage fell from 75.7% in 2019 to 58.4% in 2021, indicating potential vulnerability to measles outbreaks [41].

Figure 3. Pentavalent series coverage by dose and year. Bar chart comparing coverage rates for the 3 pentavalent doses across years.

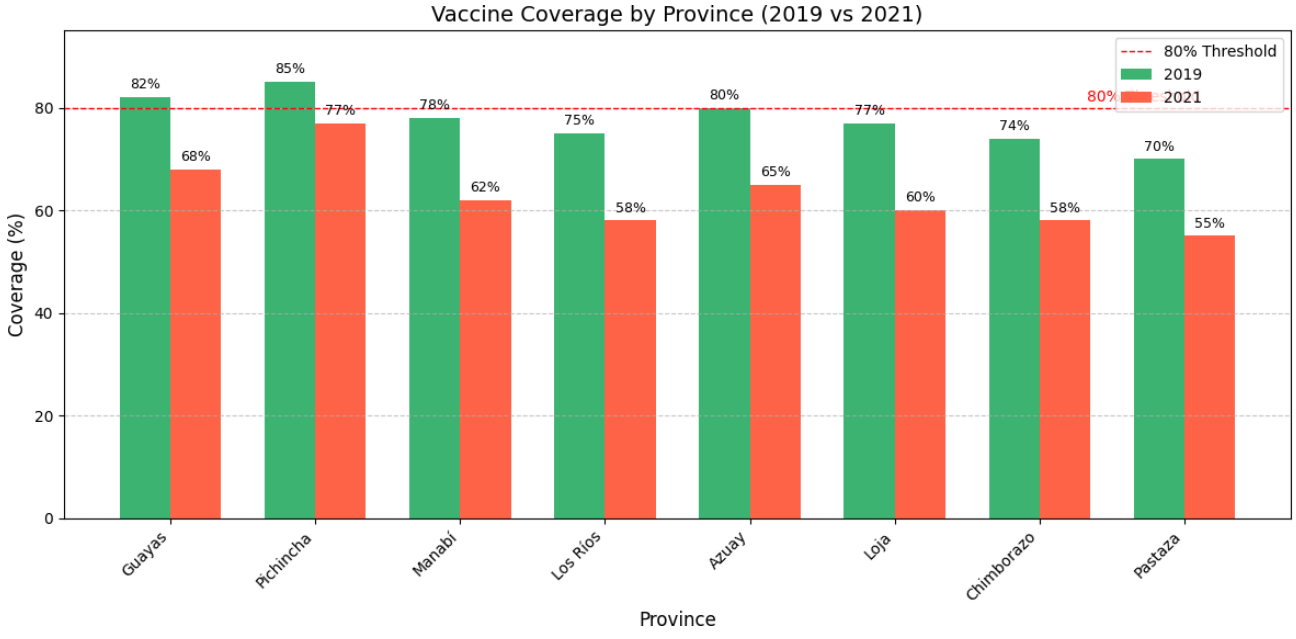


Regional and Provincial Disparities

Table 2 presents population data across Ecuador's 4 main regions and 24 provinces, providing context for understanding vaccination disparities. Analysis revealed significant geographical variations in pandemic impact [42] (Figure 4):

- Coastal region (Costa): This region, including major urban centers like Guayas (containing Guayaquil), experienced substantial coverage declines. Rural coastal provinces such as Esmeraldas and Los Ríos showed particularly severe disruptions [43].
- Highland region (Sierra): Provincial coverage varied significantly, with Pichincha (containing Quito) maintaining relatively better coverage compared to rural provinces like Bolívar and Cañar [44].
- Amazon region (Amazonía): These provinces, already facing geographical access challenges, experienced compounded difficulties during the pandemic. Remote provinces like Pastaza and Zamora Chinchipe showed marked coverage declines [45].
- Galápagos (Insular): Despite its small population, this region maintained relatively stable coverage due to its isolated nature and focused health interventions [46].

Figure 4. Regional comparison of vaccine coverage in 2019 and 2021.



Magnitude of Coverage Loss

Comparative analysis revealed that approximately 137,000 fewer vaccine doses were administered in 2020 compared to 2019, with further decreases in 2021 [47]. The pentavalent vaccine showed the most substantial absolute reduction (17.7%), followed by poliovirus, rotavirus, and pneumococcal vaccines [48].

Public Health Implications

The sustained decline in vaccination coverage through 2021 indicates that pandemic effects on childhood immunization were not temporary disruptions but represented persistent challenges to the health system [49]. Coverage rates falling below critical thresholds (particularly those below 80%) increase the risk of vaccine-preventable disease outbreaks, especially in areas with clustered susceptible populations [50].

Discussion

Principal Findings and Global Context

Our findings reveal a concerning pattern of declining routine childhood vaccination coverage in Ecuador during the COVID-19 pandemic, with sustained decreases through 2021. These results align with global trends documented worldwide, where the pandemic disrupted essential health services beyond the direct impact of SARS-CoV-2 infection [51]. The magnitude of decline in Ecuador—with some vaccines showing coverage drops exceeding 20 percentage points—represents one of the more severe impacts documented in Latin America [52].

The observed patterns are consistent with findings from other Latin American countries. Castro-Aguirre et al's [9] regional analysis of 39 countries showed significant reductions in DTP vaccine coverage in 79% of assessed regions, with Ecuador among the most affected. Our data add valuable country-specific detail to this regional picture, demonstrating the heterogeneous impact across different vaccines and geographical areas [53].

Factors Contributing to Coverage Decline

Several interconnected factors contributed to the vaccination coverage declines observed in Ecuador [54]:

- Health care system disruptions: The reallocation of health care resources to COVID-19 response efforts, including health care worker deployment to the pandemic response, reduced capacity for routine services [55]. Many health facilities were repurposed for COVID-19 care or experienced reduced operating capacity due to infection control measures [56].
- Movement restrictions and access barriers: Government-imposed lockdowns and movement restrictions, particularly strict during Ecuador's initial pandemic response, limited families' ability to access vaccination services [57]. Rural populations faced compounded challenges with transportation disruptions [58].
- Fear of infection: Parents' concerns about COVID-19 exposure in health care settings led to delayed or avoided vaccination appointments [59]. This behavioral factor persisted even as restrictions were lifted, contributing to continued coverage declines in 2021 [60].
- Supply chain disruptions: Global and regional supply chain disruptions affected vaccine availability and distribution, though specific vaccine stockout data were not consistently available for this analysis [61].

Regional Disparities and Equity Concerns

The geographical analysis revealed significant disparities in pandemic impact across Ecuador's regions [62]. Coastal and highland provinces experienced the most severe reductions, while some Amazon provinces showed variable patterns. These disparities reflect preexisting inequalities in health care access that were exacerbated during the pandemic [63].

Urban centers like Quito and Guayaquil, despite having better health care infrastructure, experienced substantial coverage declines, likely due to higher COVID-19 transmission concerns and stricter lockdown measures [64]. Rural provinces faced the dual challenge of limited health care access and additional pandemic-related barriers [65].

Indigenous and rural populations, who already faced coverage gaps before the pandemic, were disproportionately affected [66]. Arce Becerra et al's [42] study of Quito's metropolitan district showed stark urban-rural differences, with rural parish coverage declining more severely than that in urban areas.

Implications for Child Health and Disease Outbreaks

The sustained decline in vaccination coverage has serious implications for child health in Ecuador [67]. Coverage levels below 80% for most vaccines place the population at risk of vaccine-preventable disease outbreaks [68]. Of particular concern are the following:

- Measles risk: With MMR second dose coverage falling to 58.4% in 2021, Ecuador faces increased susceptibility to measles outbreaks, especially given the highly contagious nature of the measles virus and the WHO recommendation of 95% coverage for herd immunity [69].
- Pertussis and diphtheria: Declining pentavalent coverage increases the risk of these serious bacterial infections, which are particularly dangerous in young infants who rely on maternal antibodies and community immunity [70].
- Poliovirus: Although Ecuador has maintained polio-free status since 1990, reduced oral polio vaccine coverage raises concerns about potential importation and circulation of poliovirus, particularly given regional polio cases in neighboring countries [71].

Recovery Strategies and Policy Recommendations

Addressing the vaccination coverage decline requires comprehensive, multifaceted interventions [72]:

- Catch-up vaccination campaigns: Targeted mass vaccination campaigns should prioritize children who missed routine vaccinations during the pandemic. Age-appropriate catch-up schedules need implementation to ensure complete immunization, following WHO catch-up vaccination guidelines [73].

- Health system strengthening: Investment in robust health systems that can maintain essential services during emergencies is crucial. This includes adequate staffing, infrastructure improvements, and emergency preparedness protocols [74].
- Community engagement and education: Addressing vaccine hesitancy through community-based education programs, particularly targeting misinformation about COVID-19 and routine vaccines, is essential for coverage recovery [75].
- Digital health innovations: Implementation of digital vaccination registries and reminder systems can improve tracking and follow-up of children requiring catch-up vaccinations [76].
- Integrated service delivery: Combining routine vaccination with other child health services and COVID-19 vaccination efforts can improve efficiency and access [77].

Pandemic Preparedness and Resilience

Lessons learned from Ecuador's experience should inform pandemic preparedness for future health emergencies [78]:

- Essential service designation: Routine childhood vaccination should be explicitly designated as essential during health emergencies, with specific protocols to maintain service delivery [79].
- Flexible service delivery models: Developing outreach vaccination programs and mobile clinics can ensure continued access during movement restrictions [80].
- Community health workers: Training and deploying community health workers for vaccination education and basic immunization services can maintain coverage in remote areas [81].

Comparison With Global Recovery Efforts

International experience suggests that recovery of vaccination coverage requires sustained effort and multiple strategies. Countries like Rwanda and Bangladesh have demonstrated successful catch-up campaigns using innovative approaches including door-to-door vaccination and integration with COVID-19 vaccine delivery [82,83].

Study Limitations

Several limitations should be acknowledged in interpreting these findings [84]:

- Temporal scope: The analysis is limited to 2019 - 2021, preventing assessment of recovery efforts that may have begun in 2022 - 2023 [85].
- Socioeconomic data: Detailed individual-level socioeconomic data were not available, limiting the ability to fully analyze equity impacts [86].
- Causal attribution: While temporal associations are clear, directly attributing all coverage changes to COVID-19 requires careful consideration of other concurrent factors [87].
- Subnational granularity: Provincial-level analysis, while informative, may mask important local variations within provinces [88].
- Administrative versus survey data: This study relies on administrative data, which may differ from population-based survey estimates of vaccination coverage [89].

Future Research Directions

Future research should examine [90]:

- Recovery patterns in vaccination coverage post-2021
- Cost-effectiveness of different catch-up vaccination strategies
- Long-term impacts on vaccine-preventable disease incidence
- Specific interventions implemented to restore coverage
- Socioeconomic determinants of vaccination coverage disparities

Conclusions

The COVID-19 pandemic profoundly impacted routine childhood vaccination coverage in Ecuador, with declines persisting through 2021. The evidence demonstrates that, while the immediate focus on the pandemic response was necessary, the collateral damage to essential health services created new public health challenges requiring urgent attention [91].

The sustained decline in vaccination coverage—with some vaccines showing decreases exceeding 20 percentage points—places Ecuador's children at increased risk of vaccine-preventable disease outbreaks [92]. Regional disparities highlight how the pandemic exacerbated existing health inequities, with vulnerable populations facing compounded challenges in accessing immunization services [93].

Recovery requires comprehensive strategies addressing both immediate catch-up vaccination needs and longer-term health system strengthening [94]. Priority actions include implementing targeted mass vaccination campaigns, strengthening routine immunization services, and developing more resilient health systems capable of maintaining essential services during future health emergencies [95].

The findings underscore the critical importance of maintaining routine immunization programs during health crises and the need for pandemic preparedness plans that explicitly protect essential health services [96]. As Ecuador works to rebuild and strengthen its immunization program, the lessons learned from this pandemic experience must inform strategies to ensure no child is left unprotected against vaccine-preventable diseases [97].

Continued monitoring, evaluation, and research are essential to track recovery progress, evaluate intervention effectiveness, and inform evidence-based strategies for achieving and maintaining optimal vaccination coverage [98]. The protection of children's health through sustained immunization programs remains a cornerstone of public health that must be safeguarded against future disruptions [99].

Acknowledgments

This research received no external funding. The article processing charge was funded by Universidad Indoamérica.

Data Availability

The data presented in this study are available from the Ministry of Public Health of Ecuador and the National Institute of Statistics and Censuses (INEC). Data are publicly available and can be accessed through their respective official websites.

Authors' Contributions

JS contributed to the conceptualization, methodology, investigation, data curation, original draft preparation, manuscript review and editing, visualization, supervision, and project administration. KPMC and AAR Sr contributed to the validation, formal analysis, statistical analysis, and manuscript review. All authors have read and agreed to the published version of the manuscript.

Conflicts of Interest

None declared.

References

1. WHO Director-General's opening remarks at the media briefing on COVID-19. World Health Organization. 2020 Mar 11. URL: https://www.who.int/director-general/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19---11-march-2020 [accessed 2025-10-06]
2. Abbas K, Procter SR, van Zandvoort K, et al. Routine childhood immunisation during the COVID-19 pandemic in Africa: a benefit-risk analysis of health benefits versus excess risk of SARS-CoV-2 infection. Lancet Glob Health 2020 Oct;8(10):e1264-e1272. [doi: 10.1016/S2214-109X(20)30308-9] [Medline: 32687792]
3. Chandir S, Siddiqi DA, Setayesh H, Khan AJ. Impact of COVID-19 lockdown on routine immunisation in Karachi, Pakistan. Lancet Glob Health 2020 Sep;8(9):e1118-e1120. [doi: 10.1016/S2214-109X(20)30290-4] [Medline: 32615076]
4. The impact of COVID-19 on immunization services: a baseline for monitoring and evaluation. : World Health Organization; 2020 URL: https://tinyurl.com/5d4kc7m3 [accessed 2025-10-14]
5. Shet A, Carr K, Danovaro-Holliday MC, et al. Impact of the SARS-CoV-2 pandemic on routine immunisation services: evidence of disruption and recovery from 170 countries and territories. Lancet Glob Health 2022 Feb;10(2):e186-e194. [doi: 10.1016/S2214-109X(21)00512-X] [Medline: 34951973]
6. McDonald HI, Tessier E, White JM, et al. Early impact of the coronavirus disease (COVID-19) pandemic and physical distancing measures on routine childhood vaccinations in England, January to April 2020. Euro Surveill 2020 May;25(19):2000848. [doi: 10.2807/1560-7917.ES.2020.25.19.2000848] [Medline: 32431288]
7. Causey K, Fullman N, Sorensen RJD, et al. Estimating global and regional disruptions to routine childhood vaccine coverage during the COVID-19 pandemic in 2020: a modelling study. The Lancet 2021 Aug;398(10299):522-534. [doi: 10.1016/S0140-6736(21)01337-4]
8. Arsenault C, Gage A, Kim MK, et al. COVID-19 and resilience of healthcare systems in ten countries. Nat Med 2022 Jun;28(6):1314-1324. [doi: 10.1038/s41591-022-01750-1] [Medline: 35288697]
9. Castro-Aguirre IE, Alvarez D, Contreras M, et al. The impact of the coronavirus pandemic on vaccination coverage in Latin America and the Caribbean. Vaccines (Basel) 2024 Apr 25;12(5):458. [doi: 10.3390/vaccines12050458] [Medline: 38793709]
10. Esquema nacional de inmunización [Website in Spanish]. Ministerio de Salud Pública del Ecuador. 2019. URL: https://confianzaenlasvacunasla.org/wp-content/uploads/2020/11/Ecuador-ESQUEMA-DE-VACUNACION-DIC2019.pdf
11. Torres C, Sánchez I, Contreras M, García E. Disparidades en la cobertura de vacunación en Ecuador: un análisis de los determinantes sociales [Article in Spanish]. Rev Panam Salud Publica 2018;42:e123 [FREE Full text]
12. Immunization in the Americas: 2019 summary. : Pan American Health Organization; 2019 URL: https://www.paho.org/en/documents/immunization-americas-2019-summary [accessed 2025-10-14]
13. Encuesta nacional de salud y nutrición ENSANUT-ECU 2018 [Report in Spanish]. : Instituto nacional de estadística y censos del Ecuador; 2018 URL: https://anda.inec.gob.ec/anda/index.php/catalog/891 [accessed 2025-10-14]
14. Mafla-Viscarra A, Caballero E, Levy M, et al. Vaccination against COVID-19 in a geographically dispersed and underserved population, challenges and solutions in access and distribution of vaccines. F1000Res 2024;13:1294. [doi: 10.12688/f1000research.154766.1]
15. Santoli JM, Lindley MC, DeSilva MB, et al. Effects of the COVID-19 pandemic on routine pediatric vaccine ordering and administration - United States, 2020. MMWR Morb Mortal Wkly Rep 2020 May 15;69(19):591-593. [doi: 10.15585/mmwr.mm6919e2] [Medline: 32407298]
16. Roberton T, Carter ED, Chou VB, et al. Early estimates of the indirect effects of the COVID-19 pandemic on maternal and child mortality in low-income and middle-income countries: a modelling study. Lancet Glob Health 2020 Jul;8(7):e901-e908. [doi: 10.1016/S2214-109X(20)30229-1] [Medline: 32405459]

17. O'Brien KL, Baggett HC, Brooks WA, et al. Causes of severe pneumonia requiring hospital admission in children without HIV infection from Africa and Asia: the PERCH multi-country case-control study. The Lancet 2019 Aug;394(10200):757-779. [doi: 10.1016/S0140-6736(19)30721-4]
18. Ota MOC, Badur S, Romano-Mazzotti L, Friedland LR. Impact of COVID-19 pandemic on routine immunizations: evidence from the Pan American Health Organization's Regional Immunization Program. Hum Vaccin Immunother 2021;17(12):4768-4777. [doi: 10.1080/21645515.2021.1974323] [Medline: 34919493]
19. Silveira MF, Tonial CT, Goretti K. Maranhão A, et al. Missed childhood immunizations during the COVID-19 pandemic in Brazil: analyses of routine statistics and of a national household survey. Vaccine (Auckl) 2021 Jun;39(25):3404-3409. [doi: 10.1016/j.vaccine.2021.04.046]
20. von Elm E, Altman DG, Egger M, et al. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. Lancet 2007 Oct 20;370(9596):1453-1457. [doi: 10.1016/S0140-6736(07)61602-X] [Medline: 18064739]
21. Boletín de indicadores de la estrategia nacional de inmunización [Website in Spanish]. Ministerio de Salud Pública del Ecuador. 2019. URL: https://www.salud.gob.ec/boletin-de-indicadores-de-la-estrategia-nacional-de-inmunizacion/ [accessed 2025-10-14]
22. Proyecciones poblacionales y estudios demográficos [Website in Spanish]. Instituto Nacional de Estadística y Censos del Ecuador. 2019. URL: https://www.ecuadorencifras.gob.ec/proyecciones-poblacionales/ [accessed 2025-10-14]
23. Moher D, Liberati A, Tetzlaff J, Altman DG, Group P. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. PLoS Med 2009 Jul 21;6(7):e1000097. [doi: 10.1371/journal.pmed.1000097] [Medline: 19621072]
24. WHO recommendations for routine immunization - summary tables. World Health Organization. 2020. URL: https://www.who.int/teams/immunization-vaccines-and-biologicals/policies/who-recommendations-for-routine-immunization---summary-tables [accessed 2025-10-14]
25. Bravo LC, Nohynek H, Ong-Lim A, Calleja R, Ducusin MJT, Hallak J, et al. Building sustainable health systems: immunization delivery across the continuum of care. Vaccine (Auckl) 2019;37(50):7187-7194. [doi: 10.1016/j.vaccine.2019.06.087]
26. WHO position papers - recommendations for routine immunization. World Health Organization. 2019. URL: https://www.who.int/publications/m/item/table1-summary-of-who-position-papers-recommendations-for-routine-immunization [accessed 2025-10-14]
27. Brown DW, Burton A, Gacic-Dobo M, Karimov RI, Vandelaer J, Okwo-Bele JM. A summary of global routine immunization coverage through 2010. Vaccine (Auckl) 2011;29(33):5325-5331. [doi: 10.1016/j.vaccine.2011.05.035]
28. Downloading IBM SPSS Statistics 28. IBM. 2021. URL: https://www.ibm.com/support/pages/downloading-ibm-spss-statistics-28 [accessed 2025-10-14]
29. Altman DG, Bland JM. Statistics notes: the normal distribution. BMJ 1995 Feb 4;310(6975):298. [doi: 10.1136/bmj.310.6975.298] [Medline: 7866172]
30. Gardner MJ, Altman DG. Confidence intervals rather than P values: estimation rather than hypothesis testing. Br Med J (Clin Res Ed) 1986 Mar 15;292(6522):746-750. [doi: 10.1136/bmj.292.6522.746] [Medline: 3082422]
31. Fotheringham AS, Brunsdon C, Charlton M. Quantitative Geography: Perspectives on Spatial Data Analysis: SAGE Publications; 2000. URL: https://www.perlego.com/book/861255/quantitative-geography-perspectives-on-spatial-data-analysis-pdf [accessed 2025-10-14]
32. Hunter JD. Matplotlib: a 2D graphics environment. Comput Sci Eng 2007;9(3):90-95. [doi: 10.1109/MCSE.2007.55]
33. World Medical Association. World Medical Association Declaration of Helsinki: ethical principles for medical research involving human subjects. JAMA 2013;310(20):2191-2194. [doi: 10.1001/jama.2013.281053]
34. Guideline on good pharmacovigilance practices (GVP): module VIII - post-authorisation safety studies. : European Medicines Agency; 2017 URL: http://www.sefap.it/web/upload/WC500129137.pdf [accessed 2025-10-14]
35. Grimes DA, Schulz KF. Bias and causal associations in observational research. Lancet 2002 Jan 19;359(9302):248-252. [doi: 10.1016/S0140-6736(02)07451-2] [Medline: 11812579]
36. Suárez-Rodríguez GL, Salazar-Loor J, Rivas-Condo J, Rodríguez-Morales AJ, Navarro JC, Ramírez-Iglesias JR. Routine immunization programs for children during the COVID-19 pandemic in Ecuador, 2020-hidden effects, predictable consequences. Vaccines (Basel) 2022 May 27;10(6):857. [doi: 10.3390/vaccines10060857] [Medline: 35746465]
37. Velásquez-Hurtado JE, Rodríguez Y, Gonzáles M, Astete-Robilliard L, Loyola-Romaní J, Vigo WE, et al. Factors associated with routine immunization coverage in Peru's rural areas: results from a national survey. Vaccine (Auckl) 2021;39(18):2457-2464. [doi: 10.1016/j.vaccine.2021.03.064]
38. Expósito N, Martínez E, Alvarez G, Riera V, Proaño H, García S, et al. Pre-formulation study of a pentavalent DTP-HB-Hib vaccine obtained in Ecuador. Vaccimonitor 2016;25(1):25-35 [FREE Full text]
39. Wahl B, O'Brien KL, Greenbaum A, et al. Burden of Streptococcus pneumoniae and Haemophilus influenzae type b disease in children in the era of conjugate vaccines: global, regional, and national estimates for 2000-15. Lancet Glob Health 2018 Jul;6(7):e744-e757. [doi: 10.1016/S2214-109X(18)30247-X] [Medline: 29903376]

40. Troeger C, Khalil IA, Rao PC, et al. Rotavirus vaccination and the global burden of rotavirus diarrhea among children younger than 5 years. JAMA Pediatr 2018 Oct 1;172(10):958-965. [doi: 10.1001/jamapediatrics.2018.1960] [Medline: 30105384]
41. Patel MK, Goodson JL, Alexander JP Jr, et al. Progress toward regional measles elimination - worldwide, 2000-2019. MMWR Morb Mortal Wkly Rep 2020 Nov 13;69(45):1700-1705. [doi: 10.15585/mmwr.mm6945a6] [Medline: 33180759]
42. Arce Becerra CI, Zambrano Mejía LK, Nicola C. Caracterización de las zonas de riesgo susceptibles a enfermedades prevenibles por vacunación en menores de 5 años Quito-Ecuador [Article in Spanish]. Ciencia Latina 2024;8(3):5660-5676. [doi: 10.37811/cl_rcm.v8i3.11769]
43. Freeman J, Akweongo P, Zakane A, Kanmiki EW, Bangha M, Kasasa S, et al. The impact of the COVID-19 pandemic on childhood vaccination coverage in Burkina Faso. Health Policy Plan 2023;38(2):155-163. [doi: 10.1093/heapol/czac094]
44. Diaz Y, Machal AL, Lara B, Zambrano LI, Andino FG, Sierra M, et al. Surveillance of routine childhood vaccinations during the COVID-19 pandemic in Honduras. Lancet Reg Health Am 2021;3:100060. [doi: 10.1016/j.lana.2021.100060]
45. Meneses A, Burgos E, Restrepo A, Miranda K, Rueda L. Geographic barriers to routine childhood immunization in marginalized areas of Colombia: systematic review. Vaccine (Auckl) 2019;37(33):4676-4684. [doi: 10.1016/j.vaccine.2019.06.076]
46. COVID-19 situation report #50. Gavi The Vaccine Alliance. 2021. URL: https://www.gavi.org/progress-report-2021
47. The state of the world's children 2021. On my mind: promoting, protecting and caring for children's mental health. : UNICEF; 2021 URL: https://www.unicef.cn/media/20961/file/ON%20MY%20MIND%20-%20Promoting,%20protecting%20and%20caring%20for%20children%E2%80%99s%20mental%20health.pdf [accessed 2025-10-14]
48. Moraga-Llop FA, Fernández-Prada M, Grande-Tejada AM, Martínez-Alcorta LI, Moreno-Pérez D, Pérez-García C. Impacto de la pandemia por SARS-CoV-2 sobre la actividad de los centros de vacunación infantil [Article in Spanish]. An Pediatr (Barc) 2021;94(3):147-155. [doi: 10.1016/j.anpedi.2020.09.011]
49. Ahmed T, Roberton T, Vergeer P, et al. Healthcare utilization and maternal and child mortality during the COVID-19 pandemic in 18 low- and middle-income countries: an interrupted time-series analysis with mathematical modeling of administrative data. PLoS Med 2022 Aug;19(8):e1004070. [doi: 10.1371/journal.pmed.1004070] [Medline: 36040910]
50. Fine P, Eames K, Heymann DL. "Herd immunity": a rough guide. Clin Infect Dis 2011 Apr 1;52(7):911-916. [doi: 10.1093/cid/cir007] [Medline: 21427399]
51. Saxena S, Skirrow H, Bedford H, Rees J, Balasundaram V, Pollard AJ. Routine vaccination during covid-19 pandemic response. BMJ 2020 Jun 16;369:m2392. [doi: 10.1136/bmj.m2392] [Medline: 32546575]
52. Hou Z, Tong Y, Du F, et al. Assessing COVID-19 vaccine hesitancy, confidence, and public engagement: a global social listening study. J Med Internet Res 2021 Jun 11;23(6):e27632. [doi: 10.2196/27632] [Medline: 34061757]
53. Zhong Y, Clapham HE, Aishworiya R, Chua YX, Mathews J, Ong M, et al. Childhood vaccinations: hidden victims of COVID-19. Vaccine (Auckl) 2021;39(5):780-785. [doi: 10.1016/j.vaccine.2020.12.054]
54. Gaythorpe KA, Abbas K, Huber J, et al. Impact of COVID-19-related disruptions to measles, meningococcal A, and yellow fever vaccination in 10 countries. Elife 2021 Jun 24;10:e67023. [doi: 10.7554/eLife.67023] [Medline: 34165077]
55. Chen M, Lei J, Li D, Wang M, Wang H, Tian X, et al. Sustaining vaccination coverage during the COVID-19 pandemic: lessons learned from China's experiences. Vaccine (Auckl) 2022;40(13):2056-2063. [doi: 10.1016/j.vaccine.2022.02.047]
56. Siedner MJ, Kraemer JD, Meyer MJ, et al. Access to primary healthcare during lockdown measures for COVID-19 in rural South Africa: an interrupted time series analysis. BMJ Open 2020 Oct 5;10(10):e043763. [doi: 10.1136/bmjopen-2020-043763] [Medline: 33020109]
57. Cardoso P, Carvalho-Filha FSS, Silva VBM, Santos ASO, Silva AC, Figueiredo Neto JA, et al. Digital health and COVID-19: using technology to accelerate the achievement of nutrition outcomes. J Glob Health 2021;11:03085. [doi: 10.7189/jogh.11.03085]
58. Tegegne AA, Tessema GA, Kinfu Y, Padmadas SS. The immunisation of children in Ethiopia: descriptive analysis using the 2016 Ethiopian demographic and health survey data. BMC Public Health 2019;19(1):1478. [doi: 10.1186/s12889-019-7725-3]
59. Dror AA, Eisenbach N, Taiber S, et al. Vaccine hesitancy: the next challenge in the fight against COVID-19. Eur J Epidemiol 2020 Aug;35(8):775-779. [doi: 10.1007/s10654-020-00671-y] [Medline: 32785815]
60. Kiely M, Brady JE, Beil H, El-Mohandes A, El-Khorazaty MN, Perrin K. Prenatal health behaviors and birth weight among children born during the COVID-19 pandemic. JAMA Netw Open 2022;5(1):e2144984. [doi: 10.1001/jamanetworkopen.2021.44984]
61. Omer SB, Benjamin RM, Brewer NT, et al. Promoting COVID-19 vaccine acceptance: recommendations from the Lancet Commission on Vaccine Refusal, Acceptance, and Demand in the USA. The Lancet 2021 Dec;398(10317):2186-2192. [doi: 10.1016/S0140-6736(21)02507-1]
62. Bhopal S, Nielsen M. Vaccine hesitancy in low- and middle-income countries: potential implications for the COVID-19 response. Arch Dis Child 2021 Feb;106(2):113-114. [doi: 10.1136/archdischild-2020-318988] [Medline: 32912868]

63. Khubchandani J, Sharma S, Price JH, Wiblishauser MJ, Sharma M, Webb FJ. COVID-19 vaccination hesitancy in the United States: a rapid national assessment. J Community Health 2021 Apr;46(2):270-277. [doi: 10.1007/s10900-020-00958-x] [Medline: 33389421]
64. Larson HJ, Clarke RM, Jarrett C, et al. Measuring trust in vaccination: a systematic review. Hum Vaccin Immunother 2018 Jul 3;14(7):1599-1609. [doi: 10.1080/21645515.2018.1459252]
65. MacDonald NE. SAGE Working Group on Vaccine Hesitancy. Vaccine hesitancy: definition, scope and determinants. Vaccine (Auckl) 2015;33(34):4161-4164. [doi: 10.1016/j.vaccine.2015.04.036]
66. Guzman-Holst A, DeAntonio R, Prado-Cohrs D, Juliao P. Barriers to vaccination in Latin America: a systematic literature review. Vaccine (Auckl) 2020 Jan 16;38(3):470-481. [doi: 10.1016/j.vaccine.2019.10.088] [Medline: 31767469]
67. Buonsenso D, Cinicola B, Kallon MN, Iodice F. Child healthcare and immunizations in sub-Saharan Africa during the COVID-19 pandemic. Front Pediatr 2020;8:517. [doi: 10.3389/fped.2020.00517] [Medline: 32850565]
68. Guerra FM, Bolotin S, Lim G, et al. The basic reproduction number (R0) of measles: a systematic review. Lancet Infect Dis 2017 Dec;17(12):e420-e428. [doi: 10.1016/S1473-3099(17)30307-9]
69. Dabbagh A, Laws RL, Steulet C, et al. Progress toward regional measles elimination - worldwide, 2000-2017. MMWR Morb Mortal Wkly Rep 2018 Nov 30;67(47):1323-1329. [doi: 10.15585/mmwr.mm6747a6] [Medline: 30496160]
70. Klein NP, Bartlett J, Rowhani-Rahbar A, Fireman B, Baxter R. Waning protection after fifth dose of acellular pertussis vaccine in children. N Engl J Med 2012 Sep 13;367(11):1012-1019. [doi: 10.1056/NEJMoa1200850] [Medline: 22970945]
71. Polio-free Americas: 25 years and counting. : Pan American Health Organization; 2019 URL: https://www.paho.org/sites/default/files/csp30-19-e-keeping-the-region-free-of-polio_0.pdf [accessed 2025-10-14]
72. Framework for decision-making: implementation of mass vaccination campaigns in the context of COVID-19. : World Health Organization; 2020 URL: https://www.who.int/publications/i/item/WHO-2019-nCoV-Framework_Mass_Vaccination-2020.1 [accessed 2025-10-14]
73. Guidance on developing a national deployment and vaccination plan for COVID-19 vaccines. : World Health Organization; 2020 URL: https://www.who.int/publications/i/item/WHO-2019-nCoV-Vaccine_deployment-2020.1 [accessed 2025-10-14]
74. Hanson K, Brikci N, Erlangga D, et al. The Lancet Global Health Commission on financing primary health care: putting people at the centre. Lancet Glob Health 2022 May;10(5):e715-e772. [doi: 10.1016/S2214-109X(22)00005-5] [Medline: 35390342]
75. Wilson RJI, Vergélys C, Ward J, Peretti-Watel P, Verger P. Vaccine hesitancy among general practitioners in Southern France and their reluctant trust in the health authorities. Int J Qual Stud Health Well-being 2020 Dec;15(1):1757336. [doi: 10.1080/17482631.2020.1757336] [Medline: 32400299]
76. Alizadeh Khasraghi E, Marandi A, Ghazizadeh Hashemi AH, Moosavi MS, Hajimiri K, Eybpoosh S, et al. Digital health solutions in response to the first wave of COVID-19 pandemic: a comprehensive review of experiences from Iran. BMJ Innov 2021;7(4):724-732. [doi: 10.1136/bmjinnov-2021-000708]
77. Lassi ZS, Musavi NB, Maliqi B, et al. Systematic review on human resources for health interventions to improve maternal health outcomes: evidence from low- and middle-income countries. Hum Resour Health 2016 Mar 12;14:10. [doi: 10.1186/s12960-016-0106-y] [Medline: 26971317]
78. Rutter PD, Mytton OT, Mak M, Donaldson LJ. Socio-economic disparities in mortality due to pandemic influenza in England. Int J Public Health 2012 Aug;57(4):745-750. [doi: 10.1007/s00038-012-0337-1] [Medline: 22297400]
79. Maintaining essential health services: operational guidance for the COVID-19 context. : World Health Organization; 2020 URL: https://www.who.int/publications/i/item/WHO-2019-nCoV-essential_health_services-2020.2 [accessed 2025-10-14]
80. Phillips DE, Dieleman JL, Lim SS, Shearer J. Determinants of effective vaccine coverage in low and middle-income countries: a systematic review and interpretive synthesis. BMC Health Serv Res 2017 Sep 26;17(1):681. [doi: 10.1186/s12913-017-2626-0] [Medline: 28950899]
81. Cometto G, Ford N, Pfaffman-Zambruni J, et al. Health policy and system support to optimise community health worker programmes: an abridged WHO guideline. Lancet Glob Health 2018 Dec;6(12):e1397-e1404. [doi: 10.1016/S2214-109X(18)30482-0] [Medline: 30430994]
82. Okullo I, Kabbale A, Sekimpi J, Nalugoba A, Othieno E, Kiguli J, et al. COVID-19 and provision of sexual and reproductive health services in Uganda: health workers' experiences and perspectives. BMC Health Serv Res 2022;22(1):615. [doi: 10.1186/s12913-022-07978-8]
83. Khan MSI, Azad AK, Siddiquea BN, Naher S, Prioti MK, Rahaman MM, et al. The impact of COVID-19 on routine childhood immunization in Bangladesh: a mixed-method study exploring parental and health worker perspectives. BMC Public Health 2022;22(1):2236. [doi: 10.1186/s12889-022-14662-4]
84. Sanderson S, Tatt ID, Higgins JPT. Tools for assessing quality and susceptibility to bias in observational studies in epidemiology: a systematic review and annotated bibliography. Int J Epidemiol 2007 Jun;36(3):666-676. [doi: 10.1093/ije/dym018] [Medline: 17470488]
85. Bell S, Clarke R, Mounier-Jack S, Walker JL, Paterson P. Parents' and guardians' views on the acceptability of a future COVID-19 vaccine: a multi-methods study in England. Vaccine (Auckl) 2020 Nov;38(49):7789-7798. [doi: 10.1016/j.vaccine.2020.10.027]

86. Horne Z, Powell D, Hummel JE, Holyoak KJ. Countering antivaccination attitudes. Proc Natl Acad Sci U S A 2015 Aug 18;112(33):10321-10324. [doi: 10.1073/pnas.1504019112] [Medline: 26240325]
87. HILL AB. The environment and disease: association or causation? Proc R Soc Med 1965 May;58(5):295-300. [doi: 10.1177/003591576505800503] [Medline: 14283879]
88. Openshaw S. The modifiable areal unit problem. In: Concepts and Techniques in Modern Geography 1984, Vol. 38:1-41.
89. Lozano R, Soliz P, Gakidou E, et al. Benchmarking of performance of Mexican states with effective coverage. The Lancet 2006 Nov;368(9548):1729-1741. [doi: 10.1016/S0140-6736(06)69566-4] [Medline: 16488373]
90. Machingaidze S, Wiysonge CS, Hussey GD. Strengthening the expanded programme on immunization in Africa: looking beyond 2015. PLoS Med 2013;10(3):e1001405. [doi: 10.1371/journal.pmed.1001405] [Medline: 23526886]
91. Hogan DR, Stevens GA, Hosseinpoor AR, Boerma T. Monitoring universal health coverage within the Sustainable Development Goals: development and baseline data for an index of essential health services. Lancet Glob Health 2018 Feb;6(2):e152-e168. [doi: 10.1016/S2214-109X(17)30472-2] [Medline: 29248365]
92. Minta AA, Portnoy A, Karron RA, Earth G, Mvundura M, Kristiansen PA, et al. Progress on introduction and impact of meningococcal serogroup A vaccine in the meningitis belt. Clin Infect Dis 2015;61 Suppl 5:S375-S382. [doi: 10.1093/cid/civ528]
93. Wigley A, Lorin J, Hogan D, Utazi CE, Hazel E, Tatem AJ, et al. Estimates of the global burden of COVID-19 and the value of broad access to the COVID-19 vaccine. Vaccine (Auckl) 2021;39(18):2548-2557. [doi: 10.1016/j.vaccine.2021.03.013]
94. Keja K, Chan C, Hayden G, Henderson RH. Expanded programme on immunization. World Health Stat Q 1988;41(2):59-63. [Medline: 3176515]
95. GAVI Alliance strategy 2021-2025. : GAVI; 2020 URL: https://tinyurl.com/mr29pyjz [accessed 2025-10-14]
96. Cash R, Patel V. Has COVID-19 subverted global health? Lancet 2020 May 30;395(10238):1687-1688. [doi: 10.1016/S0140-6736(20)31089-8] [Medline: 32539939]
97. Goodman JL, Grabenstein JD, Braun MM. Answering key questions about COVID-19 vaccines. JAMA 2020 Nov 24;324(20):2027-2028. [doi: 10.1001/jama.2020.20590] [Medline: 33064145]
98. Immunization agenda 2030: a global strategy to leave no one behind. : World Health Organization; 2020 URL: https://www.who.int/docs/default-source/immunization/strategy/ia2030/ia2030-document-en.pdf [accessed 2025-10-14]
99. Bloom DE, Canning D, Weston M. The value of vaccination. World Econ 2005;6:15-39.

Abbreviations

DTP: diphtheria-pertussis-tetanus
INEC: National Institute of Statistics and Censuses
MMR: measles, mumps, and rubella
WHO: World Health Organization

Edited by A Grover; submitted 31.03.25; peer-reviewed by A Adekola, B Mudashiru, Z Wang; revised version received 15.05.25; accepted 19.08.25; published 17.10.25.

Please cite as:
Sanchez J, Rodriguez Sr AA, Cuello Sr KPM
Impact of the COVID-19 Pandemic on Routine Childhood Vaccination Coverage in Ecuador From 2019 to 2021: Comparative Analysis
JMIRx Med 2025;6:e75293
URL: https://xmed.jmir.org/2025/1/e75293
doi:10.2196/75293

Monte Carlo Dose Estimation of Absorbed Dose to the Hematopoietic Stem Cell Layer of the Bone Marrow Assuming Nonuniform Distribution Around the Vascular Endothelium of the Bone Marrow: Simulation and Analysis Study

Noriko Kobayashi, BA

Individual researcher, 4-16-18-1F Hamadayama Suginami-ku, Tokyo, Japan

Corresponding Author:
Noriko Kobayashi, BA
Individual researcher, 4-16-18-1F Hamadayama Suginami-ku, Tokyo, Japan

Related Articles:

Companion article: https://www.medrxiv.org/content/10.1101/2024.10.19.24315800v1

Companion article: https://med.jmirx.org/2025/1/e77775

Companion article: https://med.jmirx.org/2025/1/e77776

Companion article: https://med.jmirx.org/2025/1/e77812

Abstract

Background: Recent studies have shown that hematopoietic stem cells (HSCs) are concentrated around the endothelium of the sinusoidal capillaries. However, the current dosimetry model proposed by the International Commission on Radiological Protection (ICRP) does not account for the heterogeneity of bone marrow tissue and stem cell distribution. If the location of the hematopoietic stem cell layer differs from previous assumptions, it is necessary to re-evaluate the dose. It is especially important for short-range alpha particles where the energy deposited in the target HSC layer can vary greatly depending on the distance from the source region.

Objective: The objective of this study is to evaluate the red bone marrow doses assuming that the hematopoietic stem cell layer of the bone marrow is localized in the vascular endothelium.

Methods: A model of the trabecular bone tissues in the cervical vertebrae was developed using the Particle and Heavy Ion Transport System code. Radiation transport simulations were performed for beta and alpha radionuclides as well as noble gases, and the absorbed doses to the stem cell layer within the perivascular HSC layer of the bone marrow from inhaled radionuclides were estimated. The estimated doses were then compared with the absorbed dose based on the ICRP 60 and ICRP 103 recommendations.

Results: The absorbed doses to the bone marrow obtained from the model calculations were not significantly different from ICRP 60 and ICRP 103 for beta-nuclides. However, for alpha-nuclides, the absorbed doses were much lower than previously estimated. In addition, the contribution of red bone marrow and blood sources was greater than that of trabecular bone for alpha-nuclides. Noble gases in the red bone marrow may also affect the bone marrow stem cell layer.

Conclusions: The bone marrow dose assessment for alpha nuclides and noble gases should be re-examined using a precise model based on computed tomography images from the perspective of occupational and public radiation protection.

(JMIRx Med 2025;6:e68029) doi:10.2196/68029

KEYWORDS

stem cells; radiation; bone marrow; nuclides; noble gases

Introduction

Bone marrow is one of the most radiosensitive organs. Therefore, accurate dose assessment, considering bone microstructure and heterogeneous distribution of bone marrow tissues and cells, is critical. The International Commission on Radiological Protection (ICRP) model, currently adopted in Japan [1], assumes a homogeneous distribution of trabecular bone tissues and bone marrow stem cells.

Computational voxel phantoms have been introduced since the 2007 ICRP recommendation (ICRP 103) [2]. A precise skeletal model developed by Hough et al [3] using microcomputed tomography images of the trabecular spongiosa from an adult male cadaver has been incorporated into ICRP 133 [4]. However, hematopoietic stem cells (HSCs) are assumed to be uniformly distributed within the marrow cavities of hematopoietically active marrow [5].

Recent studies have shown that HSCs and immune cells are localized around the endothelium of bone marrow vessels [6]. One study reported that 85% of HSCs were located within 10 µm of bone marrow sinusoids in mice [7]. Kristensen et al [8] identified the microenvironment of HSCs and progenitors in the bone marrow by immunofluorescence staining of bone marrow tissue obtained from healthy volunteers. They found that the microenvironment of the HSCs is significantly enriched in sinusoids and megakaryocytes, while that of the progenitors is significantly enriched in capillaries, bone surfaces, and arteries.

Given this localized distribution of HSCs, it is necessary to re-evaluate the bone marrow dose, assuming that the HSC layer is localized around the sinusoidal capillaries of the bone marrow. This is especially important for short-range alpha particles, where the energy deposited in the target HSC layer can vary greatly depending on the distance from the source region.

Several bone marrow models have been developed for dosimetry of alpha-emitting radiopharmaceuticals, taking into account the microstructure of the bone marrow tissue. Hobbs et al [9] developed a simple geometric model of marrow cavities taking into account the distribution of bone marrow cells. They calculated the absorbed doses from ^{223}Ra in the trabecular bone surface or in the endosteal layer (layer covering the surfaces of the trabecular bone) and found that the absorbed dose was predominantly deposited near the trabecular surface and "differed markedly from a standard absorbed fraction method." Tranel et al [10] developed a cylindrical voxel bone marrow model with a blood vessel embedded in the center of the marrow and found that "the absorbed dose to the trabecular bone drops off quickly with increasing distance from the vessel wall, as the range of alphas ensures that the absorbed dose is minimal at distances greater than 100 µm." However, both studies assume a homogeneous distribution of HSCs in the bone marrow cavity. Dosimetry that accounts for the arrangement of blood vessels in the bone marrow when the source is intravascular remains a challenge.

The aim of this paper is to evaluate the bone marrow dose when HSCs are localized around sinusoidal capillaries in the bone marrow and compare it with conventional values. A geometric model of trabecular bone and bone marrow tissue was constructed at the µm scale, assuming that the HSC layer is located in the perivascular HSC layer of the sinusoids. The absorbed doses of the stem cell layer from blood and trabecular bone sources were then estimated for selected beta-nuclides, alpha-nuclides, and noble gases and compared with ICRP 60's and ICRP 103's specific absorbed fraction (SAF, fraction of radiation of energy emitted within the source region that is absorbed per mass in the target region) values. This is the first attempt at bone marrow dosimetry based on the assumption that the HSC layer is localized around sinusoidal capillaries in the bone marrow.

Methods

Geometric Modeling of Trabecular Bone and Bone Marrow Tissues

A model of the trabecular bone tissues in cervical vertebrae was created based on the data from JM-103 in the Japan Atomic Energy Agency (JAEA) Data/Code 2014 - 017 [11], using the PHITS (Particle and Heavy Ion Transport System) code version 3.17 [12]. The JM-103 data were used because a detailed weight breakdown of bone tissue and blood was not available in the ICRP 89 [13]. The cervical vertebrae were selected for modeling because they are simple in shape and easy to model.

The height of the cervical vertebrae was estimated to be 9 cm based on the following assumptions: height 171 cm, length of the spine 52 cm (about 3/10 of the height), and cervical, thoracic, and lumbar vertebrae ratio of about 2:7:3. The weight of bone tissue and blood in the cervical spine was calculated by summing the values given in the JAEA Data/Code 2014 - 017 [11]. Since the percentage of blood contained in each bone tissue was not reported, the amount of blood contained in the red bone marrow was calculated as 13.5% of the red bone marrow based on the percentages of the data reported in ICRP 89 [12] (7% of total blood for blood distributed in bone tissue and 4% for blood distributed in the red bone marrow) (Table 1). Data on the percentage of blood distributed in the sinusoids of the blood distributed in the red bone marrow were not available, so this was calculated at 89.4%, as shown in Table 2, using data from mouse bone marrow vessels by Bixel et al [14]. Material densities were set at 1.765 g/cm^3 for trabecular bone [9] and 1 g/cm^3 for red bone marrow, soft tissues, and blood.

Table . Weight of JM-103 cervical bone tissues.

Organ ID and name	Total body tissue[a] (g)	Cortical bone(g)	Trabecular bone (g)	Soft tissues (g)	Red bone marrow (g)	Blood (g)	Blood in red bone marrow[b] (g)
140 Cervical vertebra_01	0.8	__[c]	0.2	0.6	0.5	0.1	0.1
141 Cervical vertebra_02	7.1	—	3.3	3.9	2.9	0.4	0.4
142 Cervical vertebra_03	40.7	13.7	8.6	18.3	13.7	2.2	1.8
143 Cervical vertebra_04	62.5	40	—	22.5	16.9	2.8	2.3
144 Cervical vertebra_05	47.5	36.1	—	11.4	8.5	1.6	1.2
145 Cervical vertebra_06	39.5	35.6	—	4	3	—	0.4
146 Cervical vertebra_07	8.6	8.6	—	—	—	—	—
Total	206.8	134.1	12.2	60.6	45.5	7.8	6.1

[a]Total body tissue = cortical bone + trabecular bone + soft tissues.

[b]Red bone marrow 1191.6 g, blood 281.2 g: 281.2 g × 4/7/1191.6 g = 13.5%.

[c]Not available.

Table . Percentage of blood distributed in the sinusoids calculated from bone marrow vessel data of mice.

Each structure and the geometrical conditions set for the calculation.	Vessel segments (n)	Mean diameter (μm)	Cross-sectional area of blood vessels ((b/2) ^{2} × 3.14) (μm^2)[a]	Cross-sectional area of each blood vessels (c × a) (μm^2)[b]	Percentage of total cross-sectional area (%)
Arterial vessel	9	8	50.2	452.2	3.7
Postarterial capillaries	5	7.8	47.8	238.8	2.0
Intermediate capillaries	6	11.2	98.5	590.8	4.9
Sinusoidal capillaries	31	21.1	349.5	10,834.2	89.4
Total	__[c]	—	—	12,116.0	—

[a]Calculation: (mean diameter/2)2 × 3.14.

[b]Calculation: cross-sectional area of blood vessels × number of segments.

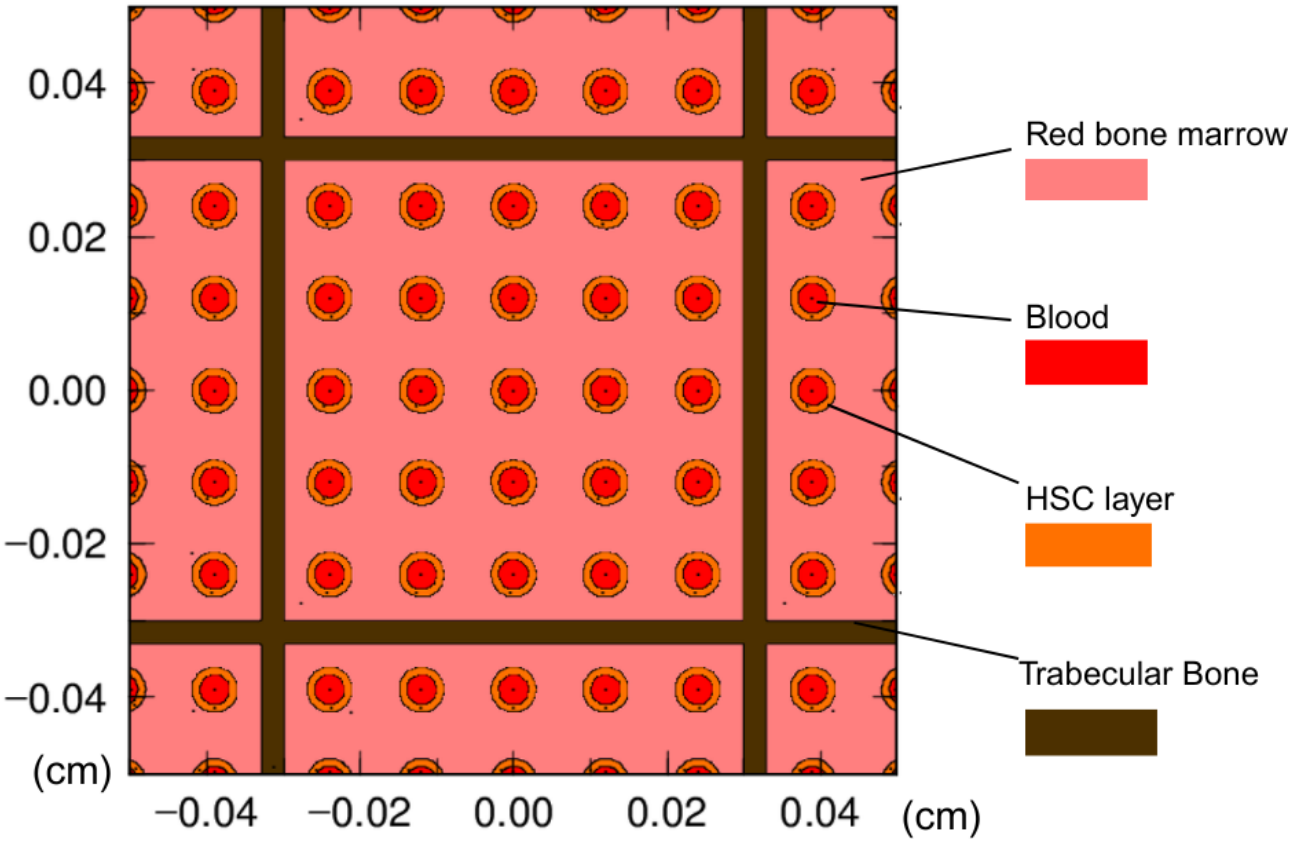
[c]Not applicable.

Based on the statement of Saladine et al [15] that sinusoids are typically 30 - 40 μm wide, the radius of the sinusoids was assumed to be 20 μm, and the number of vessels was assumed to be 40,000.

The total number of lattices was set at 1600; the internal dimension of the lattice was set at 600 μm based on the data of Parfitt et al [16]; and the external dimension of the lattice was set to 630 μm based on the weight of the trabecular bone (Figure 1).

The target part of the organ was defined as the perivascular stem cell layer 10 μm from the vascular endothelium; Acar et al [7] reported that 85% of mouse HSCs were located within 10 μm of sinusoids, and Kunisaki et al [17] reported an average distance of 14.8 μm between the vascular endothelium and HSCs. The 10 μm from the surface of the trabecular bone was defined as the trabecular surface, and the inner 30 μm was defined as the trabecular volume. Since it is impossible to model the entire trabeculae, I modeled 9 grids of 25 vessels each, for a total of 225 vessels, and multiplied the value obtained from the PHITS calculation by a factor of 40,000/225.

Figure 1. Geometry of the trabecular bone model constructed with the Particle and Heavy Ion Transport System code. HSC: hematopoietic stem cell.



Radiation Transport Simulation and Absorbed Dose Calculation

^{137}Cs, ^{131}I, and ^{90}Sr isotopes were selected for the calculation as beta-nuclides, ^{223}Ra, ^{239}Pu, ^{238}U, ^{232}Th, and ^{222}Rn as alpha-nuclides, and ^{133}Xe, ^{135}Xe, and ^{85}Kr as noble gases. Electron transport was simulated using PHITS code version 3.17 for β-radionuclides and noble gases, and alpha particles for alpha nuclides. The source regions were defined as blood, red bone marrow, trabecular bone volume, or trabecular bone based on the biokinetics of each radionuclide, and the target region was defined as the bone marrow stem cell layer 10 µm from the vascular endothelium.

For each radionuclide, electrons or alpha particles were generated in the source region, and the transferred energy distributed in the target region was calculated and converted to absorbed dose per incident particle (Gy/source). For the calculation of the beta nuclides, parameter e-type=28 was used for the source energy, which uses the DECDC [18] nuclear decay database (equivalent to ICRP 107 [19]) to obtain the energy spectra. The number of simulation trials was at least 10,000, and the statistical error in the target region was set to be less than 0.05. For the alpha radionuclide, the statistical error was set to be less than 0.5 due to the long computation time required when using the trabecular bone as a source. For ^{232}Th, the calculation was stopped with the statistical error of 0.9 because the energy distributed from the trabecular bone sources to the perivascular area was very small, which will have only a limited effect on the results and discussion even though the statistical error is relatively large. The cut-off energy for photons and electrons was set at 5 keV. Bremsstrahlung, which is a type of X-radiation emitted by charged particles when they collide or near an atomic nucleus, was included in the simulation using the Electron Gamma Shower [20] mode.

Calculation of the Number of Decays in Each Compartment

Assuming that 1 Bq (the International System of Units (SI) unit of radionuclide activity is the becquerel (Bq); 1 Bq=1 transformation/second) of radionuclide was inhaled, the number of decays in each compartment was calculated with R version 4.0.3 (R Foundation for Statistical Computing) [21] using the deSolve code [22] and the transfer coefficients presented in ICRP 56 [23], 67 [24], and 69 [25] for the current model, and those in ICRP 134 [26], 137 [27], and 141 [28] for the ICRP 103 model. The number of decays in each compartment of radionuclides transferred from the lungs to the blood was calculated for 15,800,000 minutes (10 years) for long-lived radionuclides and approximately 10,000 minutes for short-lived radionuclides. The choice of 10 years for long-lived nuclides instead of 50 years was made because of the limitations of the PC‘s performance (Intel Core i5-3337U CPU 1.8 GHz, with 7.90 GB of RAM), and 10,000 minutes for short-lived radionuclides.

For noble gases, the ICRP presents only a kinetic model for the radon dissolved in blood vessels and transported into the body. Since xenon and krypton are relatively easy to distribute in fat [29,30] as is radon [31], the transfer coefficients of radon were used for ^{133}Xe, ^{135}Xe, and ^{85}Kr. Considering that the solubility of radon in water is twice that of xenon and 4 times that of

krypton, it was assumed that 1/2 of the xenon and 1/4 of the krypton would be transferred to the blood.

The number of decays in red bone marrow blood was assumed to be proportional to the blood volume, which was 0.18% of the number of decays in whole body blood (red bone marrow blood volume in the cervical spine (volume of blood in red bone marrow of the cervical spine= 6.1g. volume of total blood in JM-103 model=3,410g. 6.1 g/3,410g=0.18%).

Calculation of the Dose Absorbed in the Perivascular Stem Cell Layer of the Bone Marrow After Inhalation of Radionuclides

Assuming that 1 L of air was inhaled after 1 hour of exposure to air containing 1 Bq/m^3 of radionuclides, the dose absorbed in the bone marrow perivascular stem cell layer was estimated by multiplying the absorbed dose determined in the section "Radiation Transport Simulation and Absorbed Dose Calculation" by the decay number calculated in the section "Calculation of the Number of Decays in Each Compartment."

Results

The absorbed doses calculated from the trabecular bone model and the comparison with the SAFs of ICRP 60 and ICRP 103 are shown in Multimedia Appendices 1-3. The absorbed dose to the perivascular HSC layer from each source was calculated for beta radionuclides (^{137}Cs, ^{131}I, and ^{90}Sr) and compared with the doses estimated using the SAF and transfer coefficients in ICRP 60 and ICRP 103, presented in Multimedia Appendix 1.

The calculation results for alpha radionuclides (^{223}Ra, ^{239}Pu, ^{238}U, ^{232}Th, and ^{222}Rn) are presented in Multimedia Appendix 2. For ^{222}Rn, only results for the PHITS trabecular bone model are shown as SAFs for radon are not provided in ICRP 60 and ICRP 103.

Results for noble gases (^{133}Xe, ^{135}Xe, and ^{85}Kr) are shown in Multimedia Appendix 3. As SAFs for noble gases are not provided in ICRP 60 and ICRP 103, only results for the PHITS trabecular bone model are shown. Table 3 summarizes the total absorbed doses to the perivascular HSC layer obtained from the PHITS calculation for each nuclide and the comparison with the ICRP 60 and ICRP 103 estimates.

Table . Summary of the calculated absorbed doses to the perivascular hematopoietic stem cell layer.

Nuclides	PHITS[a] model (Gy/source)	ICRP 60 (Gy/source)	ICRP 103 (Gy/source)	ICRP 60/PHITS	ICRP 103/PHITS
Beta-nuclides					
^{137}Cs	7.67E-09	6.83E-09	8.41E-09	0.9	1.1
^{131}I	4.26E-12	1.28E-11	4.01E-12	1.8	0.9
^{90}Sr	1.96E-08	3.43E-08	3.92E-08	1.4	2.0
Alpha-nuclides					
^{223}Ra	1.88E-10	3.92E-09	8.19E-10	20.9	4.4
^{239}Pu	1.32E-06	1.92E-05	3.23E-06	14.6	2.5
^{238}U	4.45E-10	9.70E-08	1.03E-08	217.8	23.1
^{232}Th	6.38E-07	2.26E-05	2.32E-06	35.4	3.6
^{222}Rn	1.69E-11	—[b]	—	—	—
Noble gases					
^{133}Xe	2.37E-13	—	—	—	—
^{135}Xe	3.63E-13	—	—	—	—
^{85}Kr	1.65E-13	—	—	—	—

[a]PHITS: Particle and Heavy Ion Transport System.

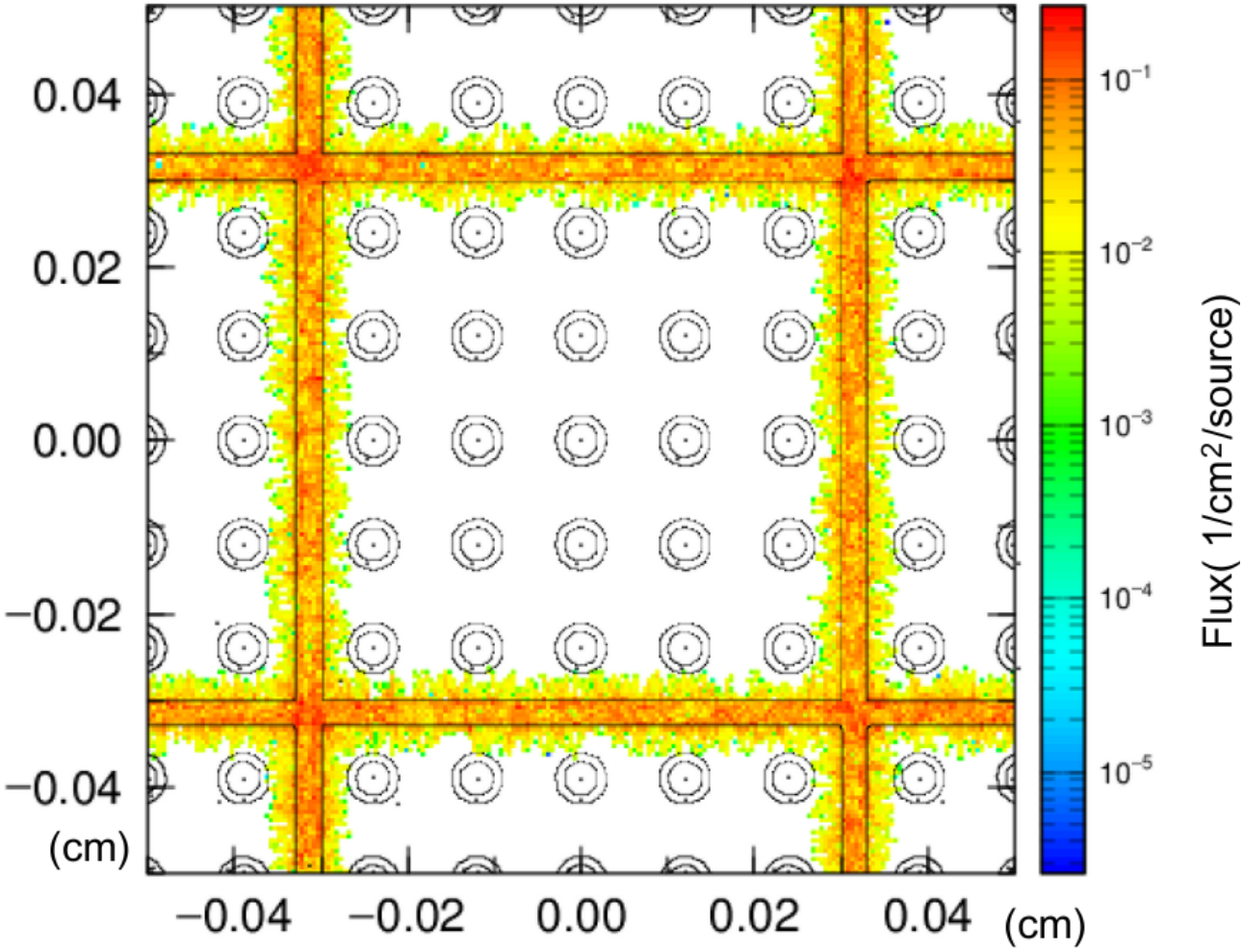
[b]Not available.

Discussion

The results show that the absorbed doses to the bone marrow obtained from the model calculations were not significantly different from ICRP 60 and ICRP 103 for beta-nuclides. In contrast, for alpha-nuclides, the absorbed doses were much lower than previously estimated. For β-nuclides, the absorbed dose per decay was higher in the PHITS model for all 3 nuclides, but the absorbed dose was almost the same as in ICRP 60 because the number of decays in each compartment changed significantly due to changes in the biokinetic model and transfer coefficients.

For the alpha nuclides, few particles reached the perivascular HSC layer from the trabecular bone source due to their short range (Figure 2, nuclide: ^{239}Pu, source organ: trabecular surface). This is consistent with the report by Tranel et al [10] that the range of alphas ensures absorbed dose is minimal at distances greater than 100 µm. Most of the dose to the perivascular HSC layer came from either the red bone marrow source or the blood source. Therefore, the dose calculated by the PHITS model was lower than the dose assessment based on the ICRP 60 recommendation, which assumes an absorbed fraction of 0.5 for the source trabecular surface and 0.05 for the trabecular bone volume. Compared to the dose estimated using the ICRP 103 SAF, the difference was smaller, about 2 to 23 times lower. Evaluation of alpha nuclides using a more accurate model is needed.

Figure 2. Particle and Heavy Ion Transport System simulation of ^{239}Pu deposited on the trabecular bone surface to the perivascular hematopoietic stem cell layer.



If HSCs are located in the perivascular HSC layer of the red bone marrow and are less susceptible to alpha radionuclides in bone sources, as suggested in Multimedia Appendix 2, the internal doses of alpha nuclides in epidemiological studies to date have been overestimated, and the actual doses to the red bone marrow may be lower. ^{223}Ra has been used for the treatment of prostate cancer, and red bone marrow doses have been evaluated by Lassmann and Nosske [32], but actual doses may be lower. As with ^{222}Rn, red bone marrow sources affect the stem cell layer. Radon biokinetics and bone marrow absorbed doses need to be evaluated, as reported by Sakoda et al [33].

Noble gases, for which internal exposure is not currently assessed, have lower absorbed doses per decay than beta and alpha nuclides, but exposure to large quantities may have radiation effects on the bone marrow stem cell layer. This may contribute to the radiation exposure of people living near accidents and nuclear power plants, where the effects of radiation exposure are controversial. A large amount of ^{133}Xe was released into the environment during the Three Mile Island accident in 1979, but the exposure of nearby residents to xenon was assessed only for external exposure; internal exposure was not included in the radiation doses. Datesman [34] noted the discrepancy between the results of physical dosimetry and biodosimetry by cytogenetic analysis of residents living near the Three Mile Island nuclear power plant. Noble gases are 10 times more soluble in lipids than in nonlipid tissues [35], and Wang et al [36] reported that bone marrow fat accounts for about 10% of total fat in healthy adults. It has also been reported that bone marrow adipocytes are located adjacent to sinusoidal blood vessels and are hematopoietic [37]. The biokinetics of xenon and other noble gases in the body and the assessment of

exposure to the bone marrow stem cell layer should be considered.

In terms of limitations, the trabecular bone model used in this paper is a simple model of part of the cervical vertebrae, although it is based on available human data. The model does not reflect differences in the mass of bone tissues according to location. The masses of bone tissues vary widely according to location in the bone, as shown in Table 4. The ratio of bone marrow and blood differs depending on the part of the bone, so the results obtained from the cervical vertebra model cannot be applied to the whole body. However, it is certainly necessary to perform a dose assessment that takes into account the fine structure of the bone and the location of the HSCs. A precise model based on microcomputed tomography images is required for dosimetry. In addition, since the transfer coefficients for noble gases are estimated from the coefficients for radon, it is necessary to construct a pharmacokinetic model based on actual measurements.

Table . Masses of bone tissues and blood of JM-103 by anatomical location.

Organ	Mass (g)						RBM/body tissue (%)	Blood/body tissue (%)
	Body tissue	RBM[a]	Cortical bone	Trabecular bone	Soft tissues	Blood		
Cranium	1346	91	774	308	264	28	6.8	2.0
Mandible	165	9	80	52	33	3	5.6	2.1
Cervical Vertebra	207	45	134	12	61	8	21.9	3.8
Thoracic Vertebra	654	187	315	77	262	32	28.7	4.9
Lumbar Vertebra	590	143	222	118	249	30	24.3	5.1
Sacrum	261	115	128	12	120	14	44.2	5.5
Clavicles	111	10	52	28	32	3	8.6	2.3
Scapulae	310	34	143	70	97	8	11.0	2.7
Sternum	107	36	39	21	47	6	33.9	5.2
Ribs	945	187	325	226	394	48	19.8	5.0
Os Coxae	1057	221	388	258	412	38	20.9	3.6
Humeri	589	29	282	122	193	10	4.9	1.7
Forearm	361	0	205	55	100	3	0.0	0.9
Wrist-Hand	220	0	115	36	69	2	0.0	1.0
Femora	1653	84	665	440	547	24	5.1	1.5
Tibiae-Fibulae-Patellae	1563	0	669	367	527	16	0.0	1.0
Ankle-Foot	872	0	299	262	313	9	0.0	1.0
Os Hyoideum	4	0	2	1	1	0	8.2	2.3
Total	11,014	1192	4837	2466	3721	281	10.8	2.6

[a]RBM: red bone marrow.

The bone marrow doses calculated with the PHITS trabecular bone marrow model, which assumes that the stem cell layer is located in the perivascular HSC layer of the sinusoids, showed that the absorbed doses from the bone marrow source and from the blood source were greater than those from trabecular bone sources for alpha nuclides. The total absorbed dose was lower than that estimated from the current ICRP models. The bone marrow dose assessments from internal exposure should be re-examined using a more detailed model of the trabecular bone marrow cavity, assuming heterogeneous distribution of HSCs and other bone marrow cells. It is also necessary to assess the effects of fat-soluble noble gases on HSCs in the bone marrow.

Acknowledgments

This paper was written by a Japanese citizen who learned the ICRP's internal radiation exposure assessment methodology after the Fukushima nuclear accident with the aim of putting an end to the controversy over radiation exposure. I would like to express my gratitude and deepest respect to Ichiro Yamaguchi, National Institute of Public Health, Japan, for his support and his efforts

in risk communication over 10 years. He answered my questions and supported the writing of this paper, but all assertions, conclusions, and mistakes herein are the sole responsibility of the author. Sincere thanks are also due to Seiko Hirota and other members of the Young Researchers Association of JHPS for allowing me to join the study group, and to the reviewer of JHPS who provided me with accurate and valuable comments. I would also like to thank Takumi Goto and Hidenaga Yoshioka for their support and assistance via the internet. I hope that the experts will address the issues raised in this paper.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Absorbed doses to the perivascular hematopoietic stem cell layer for beta radionuclides calculated with the Particle and Heavy Ion Transport System model and comparison with doses estimated using specific absorbed fraction and transfer coefficients in ICRP 60 and ICRP 103.
[DOCX File, 30 KB - xmed_v6i1e68029_app1.docx]

Multimedia Appendix 2
Absorbed doses to the perivascular hematopoietic stem cell layer for alpha radionuclides calculated with the Particle and Heavy Ion Transport System model and comparison with doses estimated using specific absorbed fraction and transfer coefficients in ICRP 60 and ICRP 103.
[DOCX File, 39 KB - xmed_v6i1e68029_app2.docx]

Multimedia Appendix 3
Absorbed doses to the perivascular hematopoietic stem cell layer for noble gases calculated with the Particle and Heavy Ion Transport System model.
[DOCX File, 19 KB - xmed_v6i1e68029_app3.docx]

References

1. International Commission on Radiological Protection. The 1990 Recommendations of the International Commission on Radiological Protection. ICRP Publication 60. Ann ICRP 1991;21(1-3):1-201. [Medline: 2053748]
2. International Commission on Radiological Protection. The 2007 Recommendations of the International Commission on Radiological Protection. ICRP Publication 103. Ann ICRP 2007;37(2-4):1-332. [doi: 10.1016/j.icrp.2007.10.003] [Medline: 18082557]
3. Hough M, Johnson P, Rajon D, et al. An image-based skeletal dosimetry model for the ICRP reference adult male--internal electron sources. Phys Med Biol 2011 Apr 21;56(8):2309-2346. [doi: 10.1088/0031-9155/56/8/001] [Medline: 21427487]
4. Bolch WE, Jokisch D, Zankl M, et al. The ICRP computational framework for internal dose assessment for reference adults: specific absorbed fractions. ICRP Publication 133. Ann ICRP 2016 Nov;45(2):5-73. [doi: 10.1177/0146645316661077] [Medline: 29749258]
5. Zankl M, Eakins J, Gómez Ros JM, et al. The ICRP recommended methods of red bone marrow dosimetry. Radiat Meas 2021 Aug;146:106611. [doi: 10.1016/j.radmeas.2021.106611]
6. Sugiyama T, Nagasawa T. Bone marrow niches for hematopoietic stem cells and immune cells. Inflamm Allergy Drug Targets 2012 Jun;11(3):201-206. [doi: 10.2174/187152812800392689] [Medline: 22452607]
7. Acar M, Kocherlakota KS, Murphy MM, et al. Deep imaging of bone marrow shows non-dividing stem cells are mainly perisinusoidal. Nature New Biol 2015 Oct 1;526(7571):126-130. [doi: 10.1038/nature15250] [Medline: 26416744]
8. Kristensen HB, Andersen TL, Patriarca A, et al. Human hematopoietic microenvironments. PLoS One 2021;16(4):e0250081. [doi: 10.1371/journal.pone.0250081] [Medline: 33878141]
9. Hobbs RF, Song H, Watchman CJ. A bone marrow toxicity model for 223Ra alpha-emitter radiopharmaceutical therapy. Phys Med Biol 2012 May 21;57(10):3207-3222. [doi: 10.1088/0031-9155/57/10/3207] [Medline: 22546715]
10. Tranel J, Feng FY, James SS, et al. Effect of microdistribution of alpha and beta-emitters in targeted radionuclide therapies on delivered absorbed dose in a GATE model of bone marrow. Phys Med Biol 2021 Jan 29;66(3):035016. [doi: 10.1088/1361-6560/abd3ef] [Medline: 33321484]
11. Manabe K, Sato K, Takahashi F. Assessment of specific absorbed fractions for photons and electrons using average adult japanese male phantom. International Nuclear Information System. 2014 Oct. URL: https://inis.iaea.org/records/nhq6n-gv229 [accessed 2025-07-03]
12. Sato T, Iwamoto Y, Hashimoto S, et al. Features of Particle and Heavy Ion Transport code System (PHITS) version 3.02. J Nucl Sci Technol 2018 Jun 3;55(6):684-690. [doi: 10.1080/00223131.2017.1419890]

13. International Commission on Radiological Protection. Basic anatomical and physiological data for use in radiological protection: reference values. Ann ICRP 2002 Sep;32(3-4):1-277. [doi: 10.1016/S0146-6453(03)00002-2] [Medline: 14506981]
14. Bixel MG, Kusumbe AP, Ramasamy SK, et al. Flow dynamics and HSPC homing in bone marrow microvessels. Cell Rep 2017 Feb 14;18(7):1804-1816. [doi: 10.1016/j.celrep.2017.01.042] [Medline: 28199850]
15. Saladine K, Sullivan S, Gan C. Human Anatomy, 5th edition: McGraw-Hill College; 2011.
16. Parfitt AM, Mathews CH, Villanueva AR, et al. Relationships between surface, volume, and thickness of iliac trabecular bone in aging and in osteoporosis. Implications for the microanatomic and cellular mechanisms of bone loss. J Clin Invest 1983 Oct;72(4):1396-1409. [doi: 10.1172/JCI111096] [Medline: 6630513]
17. Kunisaki Y, Bruns I, Scheiermann C, et al. Arteriolar niches maintain haematopoietic stem cell quiescence. Nature New Biol 2013 Oct 31;502(7473):637-643. [doi: 10.1038/nature12612] [Medline: 24107994]
18. Endo A, Yamaguchi Y. Compilation of nuclear decay data used for dose calculation:revised data for radionuclides listed in ICRP publication 38. International Nuclear Information System. 2001 Mar. URL: https://inis.iaea.org/records/evgnm-52205 [accessed 2025-07-03]
19. Eckerman K, Endo A. ICRP Publication 107. Nuclear decay data for dosimetric calculations. Ann ICRP 2008;38(3):7-96. [doi: 10.1016/j.icrp.2008.10.004] [Medline: 19285593]
20. Hirayama H, Namito Y, Bielajew AF, et al. The EGS5 code system. KEK Radiation Science Center. 2016 Jan 13. URL: https://rcwww.kek.jp/research/egs/egs5_manual/slac730-160113.pdf [accessed 2025-07-03]
21. R Core Team. R: A Language and Environment for Statistical Computing.: R Foundation for Statistical Computing; 2020. URL: https://www.r-project.org/ [accessed 2025-07-03]
22. Soetaert K, Petzoldt T, Setezer RW. Solving differential equations in R: package deSolve. J Stat Softw 2010;33(9):1-25. [doi: 10.18637/jss.v033.i09]
23. International Commission on Radiological Protection. ICRP Publication 56: age-dependent doses to members of the public from intake of radionuclides - part 1. Ann ICRP 1990;20(2):1-122. [Medline: 2633670]
24. International Commission on Radiological Protection. ICRP Publication 67: age-dependent doses to members of the public from intake of radionuclides - part 2. Ann ICRP 1990;23(3-4):1-167. [Medline: 7978694]
25. International Commission on Radiological Protection. ICRP Publication 69: age-dependent doses to members of the public from intake of radionuclides - part 3. Ann ICRP 1995;25(1):1-74. [Medline: 7486461]
26. International Commission on Radiological Protection. ICRP Publication 134: occupational intakes of radionuclides: part 2. Ann ICRP 2016;45(3-4):7-349. [Medline: 28657340]
27. International Commission on Radiological Protection. ICRP Publication 137: occupational intakes of radionuclides: part 3. Ann ICRP 2017 Dec;46(3-4):1-486. [doi: 10.1177/0146645317734963] [Medline: 29380630]
28. International Commission on Radiological Protection. ICRP Publication 141: occupational intakes of radionuclides: part 4. Ann ICRP 2019 Dec;48(2-3):9-501. [doi: 10.1177/0146645319834139] [Medline: 31850780]
29. CONN HL Jr. Equilibrium distribution of radioxenon in tissue: xenon-hemoglobin association curve. J Appl Physiol 1961 Nov;16:1065-1070. [doi: 10.1152/jappl.1961.16.6.1065] [Medline: 13880863]
30. Cohn SH, Ellis KJ, Susskind H. Evaluation of the health hazard from inhaled krypton-85. Presented at: International Symposium on Biological Implications of Radionuclides Released From Nuclear Industries; Mar 26-30, 1979; Vienna, Austria URL: https://www.osti.gov/biblio/6423958 [accessed 2025-07-03]
31. Sanjon EP, Maier A, Hinrichs A, et al. A combined experimental and theoretical study of radon solubility in fat and water. Sci Rep 2019 Jul 24;9(1):10768. [doi: 10.1038/s41598-019-47236-y] [Medline: 31341228]
32. Lassmann M, Nosske D. Dosimetry of 223Ra-chloride: dose to normal organs and tissues. Eur J Nucl Med Mol Imaging 2013 Jan;40(2):207-212. [doi: 10.1007/s00259-012-2265-y] [Medline: 23053328]
33. Sakoda A, Ishimori Y, Kawabe A, Kataoka T, Hanamoto K, Yamaoka K. Physiologically based pharmacokinetic modeling of inhaled radon to calculate absorbed doses in mice, rats, and humans. J Nucl Sci Technol 2010;47(8):731-738. [doi: 10.3327/jnst.47.731]
34. Datesman AM. Radiobiological shot noise explains Three Mile Island biodosimetry indicating nearly 1,000 mSv exposures. Sci Rep 2020 Jul 2;10(1):10933. [doi: 10.1038/s41598-020-67826-5] [Medline: 32616922]
35. Novotny JA, Parker EC, Survanshi SS, et al. Contribution of tissue lipid to long xenon residence times in muscle. J Appl Physiol (1985) 1993 May;74(5):2127-2134. [doi: 10.1152/jappl.1993.74.5.2127] [Medline: 8335539]
36. Wang H, Leng Y, Gong Y. Bone marrow fat and hematopoiesis. Front Endocrinol (Lausanne) 2018;9(694):694. [doi: 10.3389/fendo.2018.00694] [Medline: 30546345]
37. Robles H, Park S, Joens MS, et al. Characterization of the bone marrow adipocyte niche with three-dimensional electron microscopy. Bone 2019 Jan;118:89-98. [doi: 10.1016/j.bone.2018.01.020] [Medline: 29366839]

Abbreviations

HSC: hematopoietic stem cell
ICRP: International Commission on Radiological Protection

JAEA: Japan Atomic Energy Agency
PHITS: Particle and Heavy Ion Transport System
SAF: specific absorbed fraction

Edited by A Grover; submitted 26.10.24; peer-reviewed by M Gasmi, RS Gomaa Mahmoud; revised version received 21.01.25; accepted 18.05.25; published 16.07.25.

Please cite as:
Kobayashi N
Monte Carlo Dose Estimation of Absorbed Dose to the Hematopoietic Stem Cell Layer of the Bone Marrow Assuming Nonuniform Distribution Around the Vascular Endothelium of the Bone Marrow: Simulation and Analysis Study
JMIRx Med 2025;6:e68029
URL: https://xmed.jmir.org/2025/1/e68029
doi:10.2196/68029

Levels and Predictors of Knowledge, Attitudes, and Practices Regarding Contraception Among Female TV Studies Undergraduates in Nigeria: Cross-Sectional Study

Hadizah Abigail Agbo[1,2*], MBBS, MPH, MSc; Philip Adewale Adeoye[1*], MBBS, MWACP, MPH; Danjuma Ropzak Yilzung[3*], MBBS; Jawa Samson Mangut[3*], MBBS; Paul Friday Ogbada[3*], MBBS

[1]Department of Community Medicine, Jos University Teaching Hospital, Lamingo, Jos, Plateau State, Nigeria

[2]Department of Community Medicine, University of Jos, Jos, Nigeria

[3]College of Health Sciences, University of Jos, Jos, Nigeria

[*]all authors contributed equally

Corresponding Author:
Philip Adewale Adeoye, MBBS, MWACP, MPH
Department of Community Medicine, Jos University Teaching Hospital, Lamingo, Jos, Plateau State, Nigeria

Related Articles:

Companion article: https://www.medrxiv.org/content/10.1101/2024.01.01.24300698v1

Companion article: https://med.jmirx.org/2025/1/e72951

Companion article: https://med.jmirx.org/2025/1/e72949

Companion article: https://med.jmirx.org/2025/1/e72947

Abstract

Background: Access to contraception is a preventive measure against unplanned pregnancy and sexually transmitted infections; especially in sub-Saharan Africa where unmet need is a public health concern.

Objective: This study assessed the levels and predictors of knowledge, attitudes, and practices regarding contraception among female TV studies students in Nigeria.

Methods: This is a cross-sectional study conducted among female students of NTA TV College, Nigeria. Categorical sociodemographics, knowledge, attitude, and practice were presented as frequencies and proportions, while the continuous variables were presented as summary measures of central tendencies and dispersions. The primary outcome variable was the practices regarding contraception, while attitude and knowledge were secondary outcome variables, with sociodemographics as covariates. Predictors of good knowledge, attitude, and practice regarding contraception were determined by multivariable binary logistic regression, which was preceded by a bivariate regression analysis to determine candidate variables for the final model. A *P* value <.05 was determined to be statistically significant.

Results: There were 217 study participants with an average age of 22 (SD 2.6) years. Levels of good knowledge, attitude, and practice regarding contraception were reported in 55.3% (n=120), 47.5% (n=103), and 50.7% (n=110) of participants, respectively. The majority have had sex, used friends and the internet as their main sources of contraceptive information, and commonly used contraceptives such as condoms and oral contraceptive pills. The most common reason for not using contraceptives was fear of side effects or health risks. Being a young adult was a significant predictor (adjusted odds ratio [aOR] 2.6, 95% CI 1.0 - 6.7; *P*=.04) of good knowledge, while being a diploma student (aOR 2.4, 95% CI 1.2 - 4.6; *P*=.01), living off campus (aOR 2.1, 95% CI 1.0 - 4.4; *P*=.04), and good knowledge (aOR 3.8, 95% CI 2.1 - 6.9; *P*<.001) were significant predictors of good attitude. Being from the state's indigenous population (aOR 2.4, 95% CI 1.2 - 4.6; *P*=.01) and having engaged in sex (aOR 24.5, 95% CI 7.9 - 75.7; *P*<.001) were significant predictors of good contraception use.

Conclusions: Our study has shown relatively low levels of good knowledge, attitude, and practice regarding contraception and their predictors. Therefore, there is an urgent need to consistently improve advocacy, curricular development, and policies to improve knowledge, attitude, and practice regarding contraception and sexual and reproductive health services among young people.

(JMIRx Med 2025;6:e56135) doi:10.2196/56135

KEYWORDS

knowledge; attitudes; practice; contraception; regression; cross-sectional; female; students; Nigeria

Introduction

Worldwide, the proportion of women with unmet needs for modern contraception is highest in sub-Saharan Africa—twice the world's average [1]. This unmet need reportedly leads to unwanted pregnancies, unsafe abortions, and the limited ability of women to advance educationally, career-wise, and economically. The use of contraceptives should be a right-based issue that is necessary for ensuring informed choices regarding family planning. Correct use can significantly improve women's reproductive health and well-being [2]. Contraception can be an important measure against unintended pregnancy, abortion, and sexually transmitted infections (STIs), especially among young people. Recent efforts in the last decade have sought to reduce unmet needs among women and girls [3].

The proportion of youth (aged 15 - 35 years) in Nigeria is reported to be about half of the population, with about 57% who have never married [4]. This group constitutes the highest proportion of those who join the higher education system yearly. Though this group is the most sexually active and has higher contraceptive use rates, they also have the highest level of unmet needs among all population groups [5-8]. They are more likely to have premarital sex, often without protection; early, multiple, and short-lived intimate relationships; limited knowledge of sexuality issues that are required for a healthy sex life and reduction in the risk of teenage pregnancy, unsafe abortions, and STIs; and less likely to discuss family planning issues with health care providers [6,8], which poses public health and social problems in many lower- and middle-income countries, with many studies indicating increasing incidence of unsafe abortion, STIs including HIV/AIDS, violence against girls, and pregnancy-related morbidity and mortality [8-10]. Unintended and unwanted pregnancy among university students may jeopardize their academic pursuit and potential future careers.

Despite the increased accessibility of contraceptives at health facilities across the country, use remains low among young females. General uptake varies across the region, with the North Central region having the most unmet needs. Girls faced more challenges accessing contraceptives than women living with intimate partners due to the associated stigma associated with their premarital sexual activities [1,6,8,11]. Additionally, they are faced with limited access to health-related information about contraception, with regional variations, with their perception being guided by the prevailing sociocultural norms and peer influence [1,6,8]. Thus, young people lack the needed self-efficacy to negotiate a healthy sexual encounter.

Several studies have reported contraceptive knowledge, attitude, and practice among various groups of young people. Recent reports have shown that exposure to mass media communication regarding family planning increases the likelihood of use in sub-Saharan Africa [12]. With Nigeria's high fertility rate and a corresponding maternal mortality rate [6,13,14], efforts should be geared at increasing and monitoring access to family planning information and services among young women to ensure a healthy reproductive life and general well-being.

Unfortunately, there is currently a curricular deficit in the training of TV/broadcast undergraduates on health communication in Nigeria, which will continue to perpetuate professional incapacitation of future practitioners, such as limited knowledge of health issues, and the interpretation and contextualization of such to a target audience [15]. Exposure to mass media has been shown as a proven means of improving knowledge, positive attitude, advocacy, and self-efficacy for contraception use. Appropriate, adequate, and consistent training of TV producers, anchors, and writers, and other creative writers will ensure accurate and reliable health information dissemination and improve reproductive health outcomes and accountability [16].

Additionally, little is known about contraceptive knowledge, attitude, and practice among tertiary education students in mass communication, journalism, and TV studies disciplines, who may be involved in the conceptualization, development, and implementation of information, communication, and education programs, and mass media campaign activities regarding contraception among communities and the nation in the future. The state where the study was done has been shown to have the highest level of unmet needs in Nigeria's North Central region among married females [6]; therefore, there is a need to assess contraceptive behavior among the study population. This study assessed the level and predictors of contraception knowledge, attitudes, and practices of female TV studies students in Nigeria.

Methods

Overview

This is a cross-sectional study. The conceptual theory for this study is the health belief model. There are 7 constructs in the model, which were applied to the use of contraceptives among the study participants. Perceived susceptibility to unwanted pregnancy may make an individual evaluate the perceived severity or consequences of unintended pregnancy. This may drive the individual to evaluate the perceived benefits of contraceptive use to prevent the perceived threat of contraceptive nonuse. Perceived barriers to contraceptive use will be thoroughly evaluated to decide on the feasibility of contraceptive use. However, because human behavioral change takes time, there will be the need to remind individuals to adopt and maintain contraceptive use often (cues to action) via mass media, information and communication technology, family and peers, and the development of self-efficacy for contraceptive use [17].

Study Area

Plateau State is located in the North Central part of Nigeria. It covers a land area of 26,899 km^2, with an estimated population of about 4.9 million [18]. It has over 40 ethnolinguistic groups,

and each group has its distinct language. English and Hausa are common spoken languages in Plateau State. It is bounded in the northeast by Bauchi State, the northwest by Kaduna, the southwest by Nasarawa, and the southeast by Taraba State. Though situated in a tropical area, it has a near-temperate climate due to its high altitude. It has 17 local governments of which Jos North, Jos South, and Jos East make up the Jos metropolis. It has 12 higher institutions awarding various postsecondary school certificates, among which are 5 specialized educational institutions including NTA TV College [19].

NTA TV College was created in 1980 to meet the need to train the TV workforce to meet broadcast challenges in Nigeria. It is located in Rayfield, Jos South Local Government Area, Jos, Plateau State. From the beginning, it was concerned with the conduct of continuous professional development and short courses for TV industry stakeholders. It was later upgraded to offer diploma and degree programs. Nigeria's only higher institution of TV studies is currently affiliated with Ahmadu Bello University, Zaria-Nigeria, to offer a mass communication degree in TV production and journalism [20].

Study Population

The study population was female students of NTA TV College who gave their consent to participate in the study. Female students of NTA TV College who withdrew their consent to participate at any stage of the study and those who were not available for one reason or the other during the research were excluded from the study.

Sample Size Determination

Calculation of the sample size was determined using the Cochran sample size determination formula [21], mathematically expressed as:

$$n = z^2pq/d^2$$

Where n is the minimum sample size, z is the standard normal deviate at a 95% CI (1.96), p is the proportion of female undergraduates who are aware of contraception (0.84) [22], q is the alternate probability (1 – 0.84 = 0.16), and d is the precision (5%=0.05). Therefore, 206 was the estimated sample size. After adjusting for 10% nonresponse, the estimated sample size was 227.

Sampling Technique

The female students were selected using a simple random technique by balloting from the six levels during classes in the college using a proportionate approach. There were about 1021 students at the time of the study, with 220 in Ordinary National Diploma (OND) 1 (n=47, 21.7% of final sample), 208 in OND 2 (n=44, 20.3% of final sample), 29 in 100L (n=6, 2.8% of final sample), 184 in 200L (n=39, 18% of the final sample), 215 in 300L (n=46, 21.2% of the final sample), and 165 in 400L (n=35, 16.1% of final sample).

Study Instrument and Data Collection Methods

Data were collected from the female students of NTA TV College by the research team using a semistructured self-administered questionnaire after informed written consent was obtained (Multimedia Appendix 1). The questionnaire is divided into four sections: social demographic characteristics, knowledge of contraception, attitudes toward contraceptives, contraceptive practices, and sexual behavior. The questionnaire was pretested and pilot-tested among female students at the National Film Institute Jos, which has a comparable population to our study population, to identify errors, test the fidelity of the research process and the feasibility of the study, and observe the understandability of the research tools by the research participants during the pilot.

Data Management and Analysis

The data obtained were entered and analyzed using SPSS (version 25; IBM Corp). Qualitative data such as sex and religion were presented using frequencies and percentages, while quantitative data were presented using means and SDs, except when not normally distributed. Age was categorized as adolescents (≤19 years), according to the World Health Organization and published literature [8,23], and young adults (>19 years), according to the classification of the study population by Statistics Canada [24]. The average scores were used to compute the levels of contraception knowledge, attitudes, and practice, with scores less than the average classified as poor and those equal to or greater than the average score classified as good. The primary outcome variable is contraception practices, while attitude and knowledge were secondary outcome variables. Sociodemographics, knowledge, and attitude served as covariates or independent variables, as applicable. Simple logistic regression was used to determine the factor that is associated with contraception knowledge, attitudes, and practices, and to determine the candidate variables for the multivariable analysis. Variables with less than 10% probability were selected and added to the omnibus model to determine the predictors of contraception knowledge, attitudes, and practices. A P value <.05 was considered significant. Model characteristics and fitness for each multivariable logistic regression are stated with each result.

Ethical Considerations

Ethical clearance was obtained from the Jos University Teaching Hospital Research and Ethics Committee (JUTH/DCS/IREC/127/XXXI/2619). Written informed consent was obtained before participation, which was voluntary, and clients were free to withdraw consent at any point. There was no potential hazard to the study participants that might warrant exclusion or treatment. Data collection was self-administered to prevent interference by a third party when there was no other person to ensure privacy. No identification entries (name, phone numbers) were allowed. All study participants were assured that the data would be used for academic and research purposes only. No compensation was given, except for the health education on contraception after the final data collection.

Results

There was a 95.6% (217/227) response rate among study participants.

Table 1 shows the sociodemographic characteristics of the 217 participants. The majority were young adults, with an average age of 21 (range 17-32) years. The majority were either single

or separated (n=198, 91.3%), with singles being 90.8% (n=197) and separated respondents being 0.5% (n=1) of the total study population. The majority were of the Christian faith, and a little more than 10% were of the Islamic faith. Though more students were out of state compared to their ethnic origin, they were mostly in-state residents. More students were in degree programs and earned less than the minimum wage. More than two-thirds were TV journalism students, and more than three-quarters lived off campus.

Table . Sociodemographic characteristics of study participants (n=217).

Variable	Values
Age (years), mean (SD)	21.9 (2.6)
≤19 years (adolescents), n (%)	34 (15.7)
>19 years (young adults), n (%)	183 (84.3)
Marital status, n (%)	
Single/separated	198 (91.3)
Married	19 (8.7)
Religion, n (%)	
Christianity	186 (85.7)
Islam	31 (14.3)
Tribe/ethnicity, n (%)	
Plateau indigenous	91 (41.9)
Plateau nonindigenous	126 (58.1)
Program, n (%)	
Diploma	91 (41.9)
Degree	126 (58.1)
Level in school, n (%)	
OND1[a]/OND2 (early classes)	91 (41.9)
100L/200L (middle classes)	45 (20.7)
300L/400L (older classes)	81 (37.3)
Monthly income[b] (n=149), median (IQR)	17,500 (10,000- 23,000)
< 18,000, n (%)	76 (51.0)
≥ 18,000, n (%)	73 (49.0)
Home residence (n=201), n (%)	
Plateau State	126 (62.7)
Outside Plateau State	75 (37.3)
School residence (n=215), n (%)	
Campus	50 (23.3)
Off campus	165 (76.7)
Department (n=214), n (%)	
TV journalism	149 (69.6)
TV production	65 (30.4)

[a]OND: Ordinary National Diploma.

[b]A currency exchange rate of US $1= 415 is applicable as of February 18, 2022.

Table 2 shows the level of contraception knowledge, attitudes, and practices among study participants. The classification was based on the use of average scores, with the good class having at least the average score and the poor class having less than the average score. It shows that just above half reported good knowledge, attitude, and practices. The average score of contraception knowledge, attitudes, and practices were 50%, 71.3%, and 35%, respectively. Almost three-quarters (160/217, 73.7%) have had sexual intercourse.

Table 3 shows the sources of information on contraception among the participants. It shows that friends (83/236, 35.2%)

and the internet (81/236, 34.3%) were the most common sources of information on contraception (with 43/236, 18.2% using Google search and 38/236, 16.1% accessing contraceptive information via social media), which was followed by family. The least used sources were newspapers and magazines.

Table . Level of knowledge on, attitudes toward, and practices of contraception and engagement in sexual activity among study respondents (N=217).

Variables	Values
Knowledge level, n (%)	
Poor knowledge	97 (44.7)
Good knowledge	120 (55.3)
Average knowledge score (%), median (IQR)	50.0 (33.0-58.0)
Attitude level, n (%)	
Good attitude	103 (47.5)
Poor attitude	114 (52.5)
Average attitude score (%), mean (SD)	71.3 (10.6)
Practice level, n (%)	
Poor practice	107 (49.3)
Good practice	110 (50.7)
Average practice score (%), median (IQR)	35.0 (28.0-52.0)
Sexual behavior, n (%)	
Ever had sexual intercourse	160 (73.7)
Never had sexual intercourse	57 (26.3)

Table . Sources of information on contraception among study respondents (n=236).

Source[a] of contraception information	Responses, n (%)
Family	29 (12.3)
Friends	83 (35.2)
Print media newspaper	8 (3.4)
Print media magazines	7 (3.0)
Internet: Google	43 (18.2)
Internet: social media	38 (16.1)
Broadcast media TV	20 (8.5)
Health facility	8 (3.4)

[a]Participants could pick more than one source of contraceptive information, and a multiple-response analysis was done.

Table 4 shows the specific contractive methods currently being used or that were ever used among study participants. Only 85 of the 217 (39.2%) respondents disclosed the specific contraception being used. It shows that condoms (37/85, 44%) and oral contraceptive pills (OCPs; 31/85, 36%) were the most common contraceptives used by students at NTA TV College. Others, which accounted for 4.7%, have used implants, emergency contraception (EC), and other unnamed forms of contraception.

Table . Specific contraceptives currently being used or ever used among study respondents (n=85).

Variables	Responses, n (%)
Condom	37 (44)
Oral contraceptive pill	31 (36)
IUCD[a]	1 (1)
Injectable	3 (4)
Withdrawal	2 (2)
Calendar method	5 (6)
Billings	2 (2)
Others	4 (5)

[a]IUCD: intrauterine contraceptive device.

Table 5 shows the reasons why respondents do not use contraceptives. It shows that the most common reason why respondents will not use contraceptives is because of side effects, distantly followed by the perception that it increases the risk of health issues and because they are single.

Table . Reasons why respondents will not use contraceptives among study respondents (n=118)[a].

Response	Responses, n (%)
Accessibility	1 (0.8)
Based on the objection of the partner	5 (4.2)
Because I am single	11 (9.3)
Because I don’t intend on having sex	6 (5.1)
Because I don’t need it	5 (4.2)
Because of the side effect	35 (29.7)
Contraceptives sometimes are not 100% guarantee	3 (2.5)
Cultural believe	2 (1.7)
Delays pregancy	8 (6.8)
Don’t need it	7 (5.9)
I don’t know	6 (5.1)
I have used it before	2 (1.7)
I love my life	1 (0.8)
I need child	4 (3.4)
It damages the womb	5 (4.2)
It depends on your fertility	1 (0.8)
It increase the risk of health issues	12 (10.2)
It makes people gain weight	1 (0.8)
Religious belief	2 (1.7)
I calculate my fertile days	1 (0.8)

[a]Participants could pick more than one source of contraceptive information, and a multiple-response analysis was done.

Table 6 shows the predictors of contraception knowledge among study participants with the model characteristics of the multivariable logistic regression. In the bivariate analysis, being a young adult (odds ratio 3.6, 95% CI 1.6-8.0) was associated with good knowledge compared to being an adolescent in the study population. Additionally, being in the middle classes (100L/200L) was associated with good contraception knowledge compared to those in the older classes (300L/400L). The multivariable logistic regression showed that being a young adult (aged >19 years) was a significant predictor of good knowledge of contraception (adjusted odds ratio [aOR] 2.6, 95% CI 1.0-6.7; *P*=.04) compared to being an adolescent (aged ≤19 years) among the study population.

Table . Predictors of contraception knowledge among study respondents.[a]

Variable	β	OR[b] (95% CI)	*P* value	β	aOR[c] (95% CI)	*P* value
Age						
>19 years (young adults)	1.3	3.6 (1.6 - 8.0)	.002[d,e]	1.0	2.6 (1.0 - 6.7)	.04[d]
≤19 years (teenagers; reference)	—[f]	1	—	—	1	—
Marital status						
Single/separated	0.6	1.8 (0.7 - 4.6)	.23	—	—	—
Married (reference)	—	1	—	—	—	—
Religion						
Islam	0.1	1.1 (0.5 - 2.5)	.74	—	—	—
Christianity (reference)	—	1	—	—	—	—
Tribe/ethnicity						
Plateau indigenous	0.3	1.3 (0.8 - 2.3)	.31	—	—	—
Plateau nonindigenous (reference)	—	1	—	—	—	—
Program						
Degree	0.2	1.2 (0.7 - 2.1)	.52	—	—	—
Diploma (reference)	—	1	—	—	—	—
Level in school						
OND1[g]/OND2 (early classes)	0.1	1.1 (0.6 - 2.0)	.78	—	—	—
100L/200L (middle classes)	0.8	2.2 (1.0 - 4.7)	.049[d]	—	—	—
300L/400L (older classes; reference)	—	1	—	—	—	—
Monthly income[h]						
≥ 18,000	0.6	1.8 (0.9 - 3.5)	.08[d,e]	0.5	1.7 (0.9 - 3.3)	.11
< 18,000 (reference)	—	1	—	—	1	—
Home residence						
Outside Plateau State	0.1	1.0 (0.6 - 1.8)	.98	—	—	—
Plateau State (reference)	—	1	—	—	—	—
School residence						
Off campus	0.3	1.3 (0.7 - 2.5)	.97	—	—	—
Campus (reference)	—	1	—	—	—	—
Department						
TV journalism	0.1	1.0 (0.6 - 1.8)	.97	—	—	—
TV production (reference)	—	1	—	—	—	—

[a]Model characteristics: –2log likelihood 197.207, Cos & Snell R^2=0.048, Nagelkerke R^2=0.065, Hosmer-Lemeshow *P*=.22, overall percentage accuracy 60.4%.

[b]OR: odds ratio.

[c]aOR: adjusted odds ratio.

[d]Significant at *P*<.05.

[e]Candidate variables for multiple log regression at *P*<.10.

[f]Not applicable.

[g]OND: Ordinary National Diploma.

[h]A currency exchange rate of US $1= 415 is applicable as of February 18, 2022.

Table 7 shows the predictors of attitude toward contraception and the model characteristics of multivariable logistic regression. It shows that, in the bivariate analysis, being in the older and middle classes was associated with less likelihood of a good attitude toward contraception. Additionally, staying in an off-campus residence was associated with twice as higher likelihood of having a good attitude toward contraception. Good knowledge was associated with a 3.5 higher likelihood of a good attitude toward contraception among the study population. In multivariable logistic regression, being a diploma student (aOR 2.4, 95% CI 1.2-4.6; *P*=.01) was a significant predictor of a good attitude toward contraception compared to those in degree programs, having an off-campus accommodation at school was a significant predictor of good attitude (aOR 2.1, 95% CI 1.0 - 4.4; *P*=.04) compared to those with on-campus accommodations, and having good knowledge was a significant predictor of good attitude (aOR 3.8, 95% CI 2.1 - 6.9; *P*<.001) compared to those with poor knowledge.

Table . Predictors of attitude toward contraception among study respondents.[a]

Variable	β	OR[b] (95% CI)	*P* value	β	aOR[c] (95% CI)	*P* value
Age						
≤19 years (teenagers)	0.4	1.5 (0.7 - 3.1)	.29	—[d]	—	—
>19 years (young adults; reference)	—	1	—	—	—	—
Marital status						
Single/separated	0.5	1.6 (0.6 - 4.3)	.34	—	—	—
Married (reference)	—	1	—	—	—	—
Religion						
Islam	0.2	1.2 (0.6 - 2.6)	.62	—	—	—
Christianity (reference)	—	1	—	—	—	—
Tribe/ethnicity						
Plateau indigenous	0.1	1.1 (0.7 - 2.0)	.62	—	—	—
Plateau nonindigenous (reference)	—	1	—	—	—	—
Program						
Diploma	0.8	2.1 (1.2 - 3.7)	.007[e,f]	0.9	2.4 (1.2 - 4.6)	.01[e]
Degree (reference)	—	1	—	—	1	—
Level in school						
300L/400L (older classes)	–0.7	0.5 (0.3 - 0.9)	.02[e,f]	—	—	—
100L/200L (middle classes)	–0.8	0.4 (0.2 - 0.9)	.02[e,f]	–0.3	0.8 (0.3 - 1.7)	.52
OND1[g]/OND2 (early classes; reference)	—	1	—	—	1	—
Monthly income[h]						
≥ 18,000	0.5	1.6 (0.8 - 3.0)	.17	—	—	—
< 18,000 (reference)	—	1	—	—	—	—
Home residence						
Plateau State	0.1	1.1 (0.6 - 2.0)	.67	—	—	—
Outside Plateau State (reference)	—	1	—	—	—	—
School residence						
Off campus	0.7	2.1 (1.1 - 4.0)	.03[e,f]	0.8	2.1 (1.0 - 4.4)	.04[e]
Campus (reference)	—	1	—	—	1	—
Department						
TV journalism	0.1	1.1 (0.6 - 1.9)	.84	—	—	—
TV production (reference)	—	1	—	—	—	—
Level of knowledge						
Good knowledge	1.2	3.5 (2.0 - 6.1)	<.001[e,f]	1.3	3.8 (2.1 - 6.9)	<.001[e]

Variable	β	OR[b] (95% CI)	*P* value	β	aOR[c] (95% CI)	*P* value
Poor knowledge (reference)	—	1	—	—	1	—

[a]Model characteristics: –2log likelihood 264.244, Cos & Snell R^2=0.143, Nagelkerke R^2=0.191, Hosmer-Lemeshow *P*=.90, overall percentage accuracy 67.9%.

[b]OR: odds ratio.

[c]aOR: adjusted odds ratio.

[d]Not applicable.

[e]Significant at *P*<.05.

[f]Candidate variables for multiple logistic regression at *P*<.10.

[g]OND: Ordinary National Diploma.

[h]A currency exchange rate of US $1= 415 is applicable as of February 18, 2022.

Table 8 shows the predictors of contraceptive practice among the study population. In the bivariate analysis, being in the middle class was associated with a 2.5 higher likelihood of good contraceptive practice. Good knowledge of contraception was associated with a 2.5 higher likelihood of good contraceptive practice. Good attitude was associated with a twice higher likelihood of good contraceptive practice, and having engaged in sex was associated with good contraceptive practice. In the multivariable logistic regression, being from the state's indigenous (majority) population was a significant predictor of good contraceptive practice (aOR 2.4, 95% CI 1.2 - 4.6; *P*=.01) compared to those from the nonindigenous population. Having engaged in sex (aOR 24.5, 95% CI 7.9-75.7; *P*<.001) was a significant predictor of good contraceptive practice compared to those who had never engaged in sex.

Table . Predictors of the practice of contraception among study respondents.[a]

Variable	β	OR[b] (95% CI)	*P* value	β	aOR[c] (95% CI)	*P* value
Age						
>19 years (young adults)	0.7	2.1 (1.0 - 4.5)	.05[d]	–0.4	0.7 (0.2 - 2.1)	.51
≤19 years (teenagers; reference)	__[e]	1	—	—	1	—
Marital status						
Married	0.1	1.1 (0.4 - 2.8)	.86	—	—	—
Single/separated (reference)	—	1	—	—	—	—
Religion						
Christianity	0.1	1.1 (0.5 - 2.4)	.78	—	—	—
Islam (reference)	—	1	—	—	—	—
Tribe/ethnicity						
Plateau indigenous	1.0	2.7 (1.6 - 4.7)	<.001[d,f]	0.9	2.4 (1.2 - 4.6)	.01[f]
Plateau nonindigenous (reference)	—	1	—	—	1	—
Program						
Degree	0.2	1.2 (0.7 - 2.0)	.56	—	—	—
Diploma (reference)	—	1	—	—	—	—
Level in school						
OND1[g]/OND2 (early classes)	0.2	1.2 (0.6 - 2.1)	.61	—	—	—
100L/200L (middle classes)	0.9	2.5 (1.2 - 5.3)	.02[d,f]	—	—	—
300L/400L (older classes; reference)	—	1	—	—	—	—
Monthly income[h]						
≥ 18,000	0.2	1.2 (0.6 - 2.3)	.56	—	—	—
< 18,000 (reference)	—	1	—	—	—	—
Home residence						
Plateau State	0.4	1.5 (0.9 - 2.7)	.15	—	—	—
Outside Plateau State (reference)	—	1	—	—	—	—
School residence						
Campus	0.1	1.1 (0.6 - 2.0)	.83	—	—	—
Off campus (reference)	—	1	—	—	—	—
Department						
TV journalism	0.3	1.4 (0.8 - 2.5)	.26	—	—	—
TV production (reference)	—	1	—	—	—	—
Level of knowledge						
Good knowledge	0.9	2.5 (1.5 - 4.4)	.001[d,f]	0.6	1.8 (0.9 - 3.7)	.09

Variable	β	OR[b] (95% CI)	*P* value	β	aOR[c] (95% CI)	*P* value
Poor knowledge (reference)	—	1	—	—	1	—
Attitude						
Good attitude	0.7	2.1 (1.2 - 3.6)	.008[d,f]	0.5	1.6 (0.8 - 3.1)	.20
Poor attitude (reference)	—	1	—	—	1	—
Sexual behavior						
Ever	3.3	26.0 (8.9 - 75.7)	<.001[d,f]	3.2	24.5 (7.9 - 75.7)	<.001[f]
Never (reference)	—	1	—	—	1	—

[a]Model characteristics: –2log likelihood 216.338, Cos & Snell R^2=0.322, Nagelkerke R^2=0.43, Hosmer-Lemeshow *P*=.99, overall percentage accuracy 75.6%.

[b]OR: odds ratio.

[c]aOR: adjusted odds ratio.

[d]Candidate variables for multiple log regression at *P*<.10.

[e]Not applicable.

[f]Significant at *P*<.05.

[g]OND: Ordinary National Diploma.

[h]A currency exchange rate of US $1= 415 is applicable as of February 18, 2022.

Discussion

Our study shows that about half of all respondents had good knowledge, attitudes, and practices regarding contraception, with almost three-quarters having had sex and their main sources of contraceptive information being friends and the internet. Commonly used contraceptives were condoms and OCPs. A common reason for the nonuse of contraceptives was fear of side effects or health risks. Age was observed to be a significant predictor of good knowledge of contraception, while being in a diploma program (lower degree), living off campus, and having good knowledge were significant predictors of a good attitude toward contraception. Ethnicity and sexual behavior were significant predictors of good contraception use.

Our study revealed that about half of the respondents had good knowledge. This is similar to a study in Botswana [25]. However, a lower level of good knowledge was observed among students from Selangor, Malaysia; Spain; Imo State, Nigeria; and Ethiopia [26-29], while a higher level of good knowledge was observed among students in Dodoma, Tanzania; Kano and South-South, Nigeria; Pretoria, South Africa; and Kwadaso, Ghana [30-34]. These results support the evidence that one of the major professional issues in health broadcasting and programming in Nigeria is a lack of deep specialized knowledge in health communication and programming [15]. Better knowledge among this student population will improve confidence in reporting and programming, improve demand for accountability from stakeholders, increase journalist-led family planning stories and programming, and generally raise awareness in communities about issues related to contraception and general health [16].

This study also revealed that almost half of the respondents had a good attitude toward contraception. This is similar to studies in Selangor, Malaysia and the emerging region of Ethiopia [26,35]. However, higher levels of good attitude were seen in Kano, Nigeria; Adama and the emerging regions of Ethiopia; Pretoria, South Africa; Kwadaso, Ghana; and Spain [27,31,33,34,35]. Since health broadcast professionals cannot be said to be unattached, uninvolved, unbiased, and dispassionate in the production and transmission of content, their attitudes, philosophies, beliefs, and feelings might shape their approach, strategies, language choice, and angle for relaying health messages [15,16]. Counteracting negative attitudes and stereotypes should be sustained through community dialogues (while in school and during their professional life) to improve the interest of future information professionals in issues relating to family planning and general health.

It was observed that half of the study participants had good contraceptive practices. This is lower than reported among students in Kano, Nigeria [31]. This may be due to the recent 5-state public-private partnership geared toward increasing contraceptive uptake, of which Kano is a part. There was also a financially higher commitment to family planning services in these states compared to Plateau State [36,37]. A recent report of the 5-state intervention revealed increased demand generation and uptake, and improved state government financing of contraception services [36]. This government-nongovernmental organization effort might have rubbed off on young female college students in Kano. Additionally, the recent Nigerian Demographic and Health Survey reported that women of reproductive age in Kano reported a higher level of exposure to family planning messages and discussion with health care workers on family planning during their visits to health facilities compared to Plateau women of reproductive age [6]. When health communicators are also good practitioners of their message, it increases positive decision-making among target populations. Thus, public-private initiatives that engage current and future health broadcasters and program officers to improve

and sustain the current gains of contraceptive uptake should be encouraged among young people [16].

Almost three-quarters of our sample had sex. This is similar to the sexual behavior seen among students in South-South, Nigeria [32]; lower compared to students in Spain [27]; and higher than those reported among similar populations in Botswana, urban Nigerian cities, and Kilimanjaro Region of Tanzania [7,25,38,39]. This may be a result of an increased liberal worldview among young people, a sense of freedom, and a desire for sexual experimentation in the university environment. There is a need for early sexual and reproductive education to empower young people against risky sexual exploitation and behaviors.

Friends and the internet were the most common sources of information on contraception. This is similar to studies from Kilimanjaro, Tanzania; Botswana; Ilorin and South-South, Nigeria; Dodoma, Tanzania; and Spain among students [25,27,30,38,40]. However, health facilities and health care workers were the most common sources of information on contraception among similar populations in Kwadaso, Ghana and the emerging regions of Ethiopia [34,35]. Information on contraception from family members often comes out of concern that a young person is sexually or about to be sexually active, and therefore knowledge of safe sex is needed. Family members, sisters, and mothers are highly trusted based on their overall familial relationship. Additionally, trust in internet sources was often improved among young women when the source of the information is from reputable sites such as those indicating .org, .edu, and .gov [41]. Therefore, such sites should be protected from being hacked or contaminated by conspiracy theories, overt political commentaries, and unscientific content. Limited access to health information and family planning messages might be due to inadequate health broadcasting scheduling and programming. The lack of dedicated health broadcast stations and barriers created by the use of health terminologies and jargon, which might have made health messaging abstract, misunderstood, and unappreciated by the targeted public, should be addressed by relevant stakeholders in the broadcast academics, public health professionals, and the industry [15,16].

Condoms and OCPs were the most common contraceptives used among study participants. This is similar to studies from Spain; Dodoma, Tanzania; Botswana; Limpopo, South Africa; and South-South, Nigeria [27,30,39,40,42]. This may be a result of their ready availability and accessibility over the counter in many jurisdictions. The lower proportion of individuals using EC, compared to other contraceptives, in this study was similar to a recent Nigerian Demographic and Health Survey, nationally and in North-Central Nigeria, as well as a higher uptake of EC among unmarried compared to married women [6]. This is despite the high number of sexual encounters, high history of unplanned pregnancies, and a higher unmet need for contraception among this population [6,43]. There is a need for improved sensitization about ECs to stem the high level of unplanned/unwanted pregnancies and continuous risky sexual encounters. Additionally, there is a need to ensure the availability and accessibility of different contraceptive commodities through acceptable means to improve uptake among this population.

The most common reason for contraceptive nonuse was concerns about side effects and health risks among the study population. This is similar to studies among similar populations in Botswana; Pretoria, South Africa; Benin Republic; Limpopo, South Africa; and South-South, Nigeria [11,25,33,40,42]. This might have been driven by personal experiences or information received from significant others such as friends and family members or due to apparent ignorance, even when they have never used one, as seen in many low- and middle-income countries [1]. Thus, there is a need to individualize contraceptive counseling and choice when encountering young people.

This study shows that being a young adult student is a significant predictor of the acquisition of good knowledge. This is similar to national surveys from the United States and a study from the emerging regions of Ethiopia [35,44]. This may be due to less awareness among teenagers and confusing information on contraception online (to which they are more exposed than other age groups) and their limited capacity to filter and process presented information and make appropriate decisions compared to young adults [6,7,45]. Therefore, there is a need for early contraceptive information, communication, and education before the onset of sexual relations to prevent the negative consequences of unguarded sexual and reproductive behaviors.

Being a diploma student (lower degree) was a significant predictor of a good attitude toward contraception. This is converse to most studies where a higher education significantly predicted a good attitude toward contraception [1,35]. Our result may be due to increased information fatigue following information overload that may occur, which might reduce risk perception and increase nonchalant attitude toward contraception. Information received might have been contaminated over the years with disinformation and misinformation to which higher-level degree students might have been exposed to over the years. Thus, information managers and regulators should ensure that the information provided is of high value and an opportunity for updates targeted at a specific population without infringing on human rights. Additionally establishing a consistent, stand-alone, and well-grounded health broadcasting curriculum in the schools of TV studies, journalism, and mass communication might produce an improved attitude toward contraception and help filter out disinformation during undergraduate years, which might be a departure from the current state of health broadcast training in Nigeria [15,16].

This study shows that being off campus was a significant predictor of a good attitude toward contraception. Disaggregation of the study data based on school residence shows that off-campus students were older students, were married, and had higher income and a higher level of knowledge about contraception compared to on-campus students. Similar demography of off-campus students has been reported among undergraduates in the United States [46]. These demographic characteristics are significant predictors of a good attitude toward contraception [35]. Thus, overly restrictive policies on contraceptive access and stigmatization by on-campus health care providers should be addressed to improve contraceptive uptake as needed.

Good knowledge was shown to be a good predictor of a good attitude toward contraception. This is similar to studies from emerging regions in Ethiopia and Botswana [25,35]. Consistent, appropriate, and targeted delivery of contraceptive information, education, and communication through media advocacy will improve the attitude of current and future health broadcasters, editors, and programmers, which will aid their confidence in delivering family planning messaging and programming and further improve contraceptive use among young people and the population in general [16].

Being Plateau indigenous (a conglomeration of about 50 ethnic groups and a majority population in the state) was a significant predictor of contraceptive use. This is similar to studies from similar populations from South-South, Nigeria; Selangor, Malaysia; and the United States where being a member of the majority population is a significant predictor of contraceptive use [26,40,47]. This might be due to disparities in the levels of awareness, knowledge, attitude, and access related to contraception that have been reported in the majority population. In some instances, minority populations may be wary of the government's intention to limit minority populations and be skeptical about the safety of government-sanctioned contraceptives [47,48]. The minority ethnic differences especially in minority populations should spur health care providers to provide necessary contraceptive education. Innovative counseling approaches could improve women's ability to make informed decisions. Interventions to reach out to tertiary students, especially those from minority backgrounds, should be instituted in schools to provide information, communication, and education opportunities to male and female students. Since none of the respondents mentioned any reference to sex education in schools [49], there may be a need to review the impact of the current sex education policies in schools on the sexual and reproductive health behavior of young people.

This study shows that being involved in sexual relations is a significant predictor of good contraception practice. This is similar to a study from Kilimanjaro, Tanzania among a similar population [38]. Sexually active individuals see the need to prevent unwanted pregnancy, STIs, and pregnancy-related health risks as they delay marriage to complete education while pursuing sexual relationships [9,10,50]. It also helps girls and women achieve empowerment to live a healthy and economically productive life. Studies have shown that sexually inactive young people often cite their sexual inactivity as a reason for the nonuse of contraception [42]. The health system should, therefore, be well prepared to assist young people who might need a full range of sexual and reproductive health services whenever and wherever they need them without experiencing any form of hardship.

First, the outcome of this study cannot be generalized to all universities, but the status of the college being the only TV studies college in Africa can provide insight into the knowledge, attitudes, and practices related to contraception among future TV professionals. Second, there may also be a social desirability bias as respondents might have underreported sexual behaviors and contraceptive use. This was minimized by ensuring confidentiality, anonymity, and privacy during and after the study. Third, the study is gender biased, as the study population comprised females only. Though male involvement is a great goal in achieving optimal sexual and reproductive health, females experience more reproductive health issues, have less autonomy over life choices and decisions, are exposed to more stigma while accessing contraceptive commodities, have a higher incidence of STIs, and have a higher incidence of child marriages compared to their male counterparts of similar age. All these adverse inequalities cause young females to experience higher consequential adverse outcomes, especially in low- and middle-income countries, and the need for more studies on their sexual and reproductive health [8].

Female undergraduate NTA TV College students in this study had relatively low levels of good knowledge, attitudes, and practices related to contraception. There is a need for an appropriate and consistent awareness campaign via acceptable media and curricular improvement among TV studies undergraduate students to improve their current knowledge, attitudes, and utilization of contraception. Parent-child communication should be encouraged and supported to improve contraceptive knowledge, attitude, and practice, as the family is the first educational institution in the life of children as they prepare to face the world. There is also a need to evaluate and improve the current comprehensive sex education in many sub-Saharan African countries to have more robust training for young people on sexual and reproductive health as early as possible. There is an urgent need to reform current advocacy efforts, sexual and reproductive health services, and policies to improve contraceptive knowledge, attitudes, and use among young people.

Acknowledgments

We are grateful to the students of NTA TV College of Nigeria for their willingness to participate in the study.

Authors' Contributions

Conceptualization and design: HAA, PAA, DRY, JSM, PFO
Data acquisition: DRY, JSM, PFO
Data analysis and interpretation: PAA
Drafting and critical review: HAA, PAA
Final approval: HAA, PAA, DRY, JSM, PFO

Conflicts of Interest

None declared.

Multimedia Appendix 1
Questionnaire.
[DOCX File, 20 KB - xmed_v6i1e56135_app1.docx]

References

1. Sedgh G, Ashford LS, Hussain R. Unmet need for contraception in developing countries: examining women's reasons for not using a method. 2016 Jun. URL: https://www.guttmacher.org/sites/default/files/report_pdf/unmet-need-for-contraception-in-developing-countries-report.pdf [accessed 2025-03-28]
2. Hardee K, Jordan S. Advancing rights-based family planning from 2020 to 2030. Open Access J Contracept 2021 Sep 9;12:157-171. [doi: 10.2147/OAJC.S324678] [Medline: 34531690]
3. Cohen SA. London summit puts family planning back on the agenda, offers new lease on life for millions of women and girls. Guttmacher Institute. 2012. URL: https://www.guttmacher.org/sites/default/files/article_files/gpr150320.pdf [accessed 2022-03-15]
4. National baseline youth survey: final report. National Bureau of Statistics. 2012. URL: https://www.nigerianstat.gov.ng/pdfuploads/2102%20National%20Baseline%20Youth%20Survey%20Report_1.pdf [accessed 2022-03-15]
5. Mutsindikwa T, Ashipala DO, Tomas N, Endjala T. Knowledge, attitudes and practices of contraception among tertiary students at the university campus in Namibia. Global J Health Sci 2019;11(6):180. [doi: 10.5539/gjhs.v11n6p180]
6. National Population Commission. Nigeria demographic and health survey 2018. The DHS Program. 2019 Oct. URL: https://dhsprogram.com/pubs/pdf/FR359/FR359.pdf [accessed 2025-03-28]
7. Bajoga UA, Atagame KL, Okigbo CC. Media influence on sexual activity and contraceptive use: a cross sectional survey among young women in urban Nigeria. Afr J Reprod Health 2015 Sep;19(3):100-110. [Medline: 26897918]
8. Liang M, Simelane S, Fortuny Fillo G, et al. The state of adolescent sexual and reproductive health. J Adolesc Health 2019 Dec;65(6S):S3-S15. [doi: 10.1016/j.jadohealth.2019.09.015] [Medline: 31761002]
9. Grant C. Benefits of investing in family planning. GOV.UK. 2016 Dec 1. URL: https://assets.publishing.service.gov.uk/media/5b97f5f940f0b6789a513262/021_Benefits_of_investing_in_family_planning__K4D_template_.pdf [accessed 2025-03-28]
10. Singh S, Remez L, Sedgh G, Kwok L, Onda T. Abortion worldwide 2017: uneven progress and unequal access. Guttmacher. 2018. URL: https://www.guttmacher.org/sites/default/files/report_pdf/abortion-worldwide-2017.pdf [accessed 2025-03-28]
11. Ahissou NCA, Benova L, Delvaux T, et al. Modern contraceptive use among adolescent girls and young women in Benin: a mixed-methods study. BMJ Open 2022 Jan 4;12(1):e054188. [doi: 10.1136/bmjopen-2021-054188] [Medline: 34983766]
12. Babalola S, Figueroa ME, Krenn S. Association of mass media communication with contraceptive use in sub-Saharan Africa: a meta-analysis of demographic and health surveys. J Health Commun 2017 Nov;22(11):885-895. [doi: 10.1080/10810730.2017.1373874] [Medline: 29125805]
13. World Health Organization. Trends in maternal mortality: 1990-2015: estimates from WHO, UNICEF, UNFPA, World Bank Group and the United Nations Population Division. ReliefWeb. 2015. URL: https://reliefweb.int/sites/reliefweb.int/files/resources/9789241565141_eng.pdf [accessed 2025-03-28]
14. Ope BW. Reducing maternal mortality in Nigeria: addressing maternal health services' perception and experience. J Global Health Rep 2020 May 18;4:e2020028. [doi: 10.29392/001c.12733]
15. Obong UA, Senam N. Issues in health broadcasting in Nigeria. Recent Arch Journalism Mass Commun 2024 Aug 13;1(3):555562. [doi: 10.19080/RAJMC.2024.01.555562]
16. Choge I, Mwalimu R, Mulyanga S, Njiri S, Kwachi B, Ontiri S. Media advocacy in catalyzing actions by decision-makers: case study of the advance family planning initiative in Kenya. Front Glob Womens Health 2023 Jun 4;4:1168297. [doi: 10.3389/fgwh.2023.1168297] [Medline: 37346972]
17. Jones CL, Jensen JD, Scherr CL, Brown NR, Christy K, Weaver J. The Health Belief Model as an explanatory framework in communication research: exploring parallel, serial, and moderated mediation. Health Commun 2015;30(6):566-576. [doi: 10.1080/10410236.2013.873363] [Medline: 25010519]
18. Plateau · population. Population City. 2022 Mar 7. URL: http://population.city/nigeria/adm/plateau/ [accessed 2025-03-28]
19. Samuel Okwa. Employment and Labour Market Analysis - Plateau State [Internet]. 2022. URL: https://psosic.org/docs/Plateau State Labour Information.pdf [accessed 2025-04-18]
20. About us. NTA Television College. 2021 Mar 7. URL: https://ntatvc.edu.ng/?page_id=2876 [accessed 2025-03-28]
21. Bolarinwa OA. Sample size estimation for health and social science researchers: the principles and considerations for different study designs. Niger Postgrad Med J 2020;27(2):67-75. [doi: 10.4103/npmj.npmj_19_20] [Medline: 32295935]
22. Eniojukan JF. Knowledge, perception and practice of contraception among staff and students in a university community in Delta State, Nigeria. Pharm Biosciences J 2016 Feb;4(1):71-81. [doi: 10.20510/ukjpb/4/i1/87848]

23. Adolescent health. World Health Organization. 2024 Jun 29. URL: https://www.who.int/health-topics/adolescent-health [accessed 2025-03-28]
24. Infographic 5: Plateau in the share of young adults living with their parents from 2016 to 2021. Statistics Canada. 2022 Jun 29. URL: https://www150.statcan.gc.ca/n1/daily-quotidien/220713/g-a005-eng.htm [accessed 2025-03-28]
25. Kgosiemang B, Blitz J. Emergency contraceptive knowledge, attitudes and practices among female students at the University of Botswana: a descriptive survey. Afr J Prim Health Care Fam Med 2018 Sep 6;10(1):e1-e6. [doi: 10.4102/phcfm.v10i1.1674] [Medline: 30198288]
26. Oo MS, Ismail NBM, Ean WR, Hamid HA, Affendi NR. Knowledge, attitude and perception of contraception among medical students in Universiti Putra Malaysia. Malaysian J Public Health Med 2019 Apr 1;19(2):11-19. [doi: 10.37268/mjphm/vol.19/no.2/art.165]
27. Sanz-Martos S, López-Medina IM, Álvarez-García C, et al. Young nursing student's knowledge and attitudes about contraceptive methods. Int J Environ Res Public Health 2020 Aug 13;17(16):5869. [doi: 10.3390/ijerph17165869] [Medline: 32823694]
28. Arisukwu O, Igbolekwu CO, Efugha I, Nwogu JN, Osueke NO, Oyeyipo E. Knowledge and perception of emergency contraceptives among adolescent girls in Imo State, Nigeria. Sexuality Cult 2019 Aug 26;24:273-290. [doi: 10.1007/s12119-019-09639-x]
29. Fikre R, Amare B, Tamiso A, Alemayehu A. Determinant of emergency contraceptive practice among female university students in Ethiopia: systematic review and meta-analysis. Contracept Reprod Med 2020 Oct 5;5:18. [doi: 10.1186/s40834-020-00123-8] [Medline: 33029382]
30. Kara WSK, Benedicto M, Mao J. Knowledge, attitude, and practice of contraception methods among female undergraduates in Dodoma, Tanzania. Cureus 2019 Apr 2;11(4):e4362. [doi: 10.7759/cureus.4362] [Medline: 31192067]
31. Gajida AU, Takai IU, Haruna IU, Bako KA. Knowledge, attitude and practice of modern contraception among women of reproductive age in urban area of Kano, North-Western Nigeria. J Med Trop 2019;21(2):67-72. [doi: 10.4103/jomt.jomt_9_19]
32. Duru CB, Nnebue CC, Uwakwe KA, et al. Sexual behaviours and contraceptive use among female secondary school adolescents in a rural town in RiverS state, South-south Nigeria. Niger J Med 2015;24(1):17-27. [doi: 10.12691/ajmsm-3-5-1] [Medline: 25807669]
33. Bongongo T, Govender I. Knowledge, attitudes and practices of contraceptive methods among women seeking voluntary termination of pregnancy at Jubilee Hospital, Pretoria, South Africa. Afr J Prim Health Care Fam Med 2019 Aug 15;11(1):e1-e5. [doi: 10.4102/phcfm.v11i1.1919] [Medline: 31478741]
34. Yeboah DS, Appiah MA, Kampitib GB. Factors influencing the use of emergency contraceptives among reproductive age women in the Kwadaso Municipality, Ghana. PLoS One 2022 Mar 3;17(3):e0264619. [doi: 10.1371/journal.pone.0264619] [Medline: 35239714]
35. Bekele D, Surur F, Nigatu B, et al. Knowledge and attitude towards family planning among women of reproductive age in emerging regions of Ethiopia. J Multidiscip Healthc 2020 Nov 4;13:1463-1474. [doi: 10.2147/JMDH.S277896] [Medline: 33177832]
36. Desmon S. Putting five Nigerian states in driver's seat on family planning. Johns Hopkins Center for Communication Programs. 2021 Aug 23. URL: https://ccp.jhu.edu/2021/08/23/family-planning-nigerian-tci-success/ [accessed 2025-03-28]
37. Nigeria's Plateau State releases $25,125 for family planning and reproductive health in the 2016 state health budget. Advance Family Planning. 2017 Mar 17. URL: https://www.advancefamilyplanning.org/nigerias-plateau-state-releases-25125-family-planning-and-reproductive-health-2016-state-health [accessed 2025-03-28]
38. Sweya MN, Msuya SE, Mahande MJ, Manongi R. Contraceptive knowledge, sexual behavior, and factors associated with contraceptive use among female undergraduate university students in Kilimanjaro region in Tanzania. Adolesc Health Med Ther 2016 Oct 3;7:109-115. [doi: 10.2147/AHMT.S108531] [Medline: 27757057]
39. Hoque ME, Ntsipe T, Mokgatle-Nthabu M. Awareness and practices of contraceptive use among university students in Botswana. SAHARA J 2013;10(2):83-88. [doi: 10.1080/17290376.2013.869649] [Medline: 24405283]
40. Agbo OJ, Eguvbe AO, Alabra PW, Alagoa DO. Knowledge of modern contraceptives methods and its uptake among female students of a tertiary educational institution in South- South Nigeria. Eur J Med Health Sci 2020;2(5). [doi: 10.24018/ejmed.2020.2.5.450]
41. Freeman JL, Caldwell PHY, Bennett PA, Scott KM. How adolescents search for and appraise online health information: a systematic review. J Pediatr 2018 Apr;195:244-255. [doi: 10.1016/j.jpeds.2017.11.031] [Medline: 29398062]
42. Raselekoane RN, Chinyakata R, Gwatimba L. An investigation of the knowledge, attitudes and use of contraceptives by youth development female students at the University of Venda. J Hum Ecol 2018;61(1–3):1-8. [doi: 10.1080/09709274.2018.1444452]
43. Adeoye PA, Adeniji T, Agbo HA. Predictors of good contraception attitude and practice among female students of television studies in Nigeria: a secondary analysis. medRxiv. Preprint posted online on Aug 4, 2024. [doi: 10.1101/2024.02.26.24303367]
44. Craig AD, Dehlendorf C, Borrero S, Harper CC, Rocca CH. Exploring young adults' contraceptive knowledge and attitudes: disparities by race/ethnicity and age. Womens Health Issues 2014;24(3):e281-e289. [doi: 10.1016/j.whi.2014.02.003] [Medline: 24725755]

45. Grootens-Wiegers P, Hein IM, van den Broek JM, de Vries MC. Medical decision-making in children and adolescents: developmental and neuroscientific aspects. BMC Pediatr 2017 May 8;17(1):120. [doi: 10.1186/s12887-017-0869-x] [Medline: 28482854]
46. Blagg K, Rosenboom V. Who lives off campus? An analysis of living expenses among off-campus undergraduates. Urban Institute. 2017 Oct. URL: https://www.urban.org/sites/default/files/publication/94016/who-lives-off-campus.pdf [accessed 2025-03-28]
47. Dehlendorf C, Park SY, Emeremni CA, Comer D, Vincett K, Borrero S. Racial/ethnic disparities in contraceptive use: variation by age and women's reproductive experiences. Am J Obstet Gynecol 2014 Jun;210(6):526. [doi: 10.1016/j.ajog.2014.01.037] [Medline: 24495671]
48. Payne C, Fanarjian N. Seeking causes for race-related disparities in contraceptive use. Virtual Mentor 2014 Oct 1;16(10):805-809. [doi: 10.1001/virtualmentor.2014.16.10.jdsc1-1410] [Medline: 25310047]
49. Wekesah FM, Nyakangi V, Onguss M, Njagi J, Bangha M. Comprehensive sexuality education in sub-Saharan Africa. African Population and Health Research Center. 2019. URL: https://aphrc.org/wp-content/uploads/2019/12/COMPREHENSIVE-SEXUALITY-EDUCATION-IN-SUB-SAHARAN-AFRICA-1.pdf [accessed 2025-03-28]
50. Aina IT, Aina-Pelemo AD. The use of contraceptives in Nigeria: benefits, challenges and probable solutions. J Law Policy Globalization 2019 Jun 30;86:88-99. [doi: 10.7176/JLPG/86-09]

Abbreviations

aOR: adjusted odds ratio
EC: emergency contraception
OCP: oral contraceptive pill
OND: Ordinary National Diploma
STI: sexually transmitted infection

Edited by A Schwartz, E Meinert; submitted 07.01.24; peer-reviewed by B Nwankwo, K Biswas; revised version received 05.10.24; accepted 13.12.24; published 08.05.25.

Please cite as:
Agbo HA, Adeoye PA, Yilzung DR, Mangut JS, Ogbada PF
Levels and Predictors of Knowledge, Attitudes, and Practices Regarding Contraception Among Female TV Studies Undergraduates in Nigeria: Cross-Sectional Study
JMIRx Med 2025;6:e56135
URL: https://xmed.jmir.org/2025/1/e56135
doi:10.2196/56135

Large Language Models for Pediatric Differential Diagnoses in Rural Health Care: Multicenter Retrospective Cohort Study Comparing GPT-3 With Pediatrician Performance

Masab Mansoor[1], BS, MBA, DBA; Andrew F Ibrahim[2], BS; David Grindem[3], DO; Asad Baig[4], MD

[1]Edward Via College of Osteopathic Medicine, 4408 Bon Aire Dr, Monroe, LA, United States

[2]Texas Tech University Health Sciences Center School of Medicine, Lubbock, TX, United States

[3]Mayo Clinic, Rochester, MN, United States

[4]Department of Radiology, Columbia University Medical Center, New York, NY, United States

Corresponding Author:
Masab Mansoor, BS, MBA, DBA
Edward Via College of Osteopathic Medicine, 4408 Bon Aire Dr, Monroe, LA, United States

Related Articles:

Companion article: https://www.medrxiv.org/content/10.1101/2024.08.09.24311777v1

Companion article: https://med.jmirx.org/2025/1/e73264

Companion article: https://med.jmirx.org/2025/1/e73258

Abstract

Background: Rural health care providers face unique challenges such as limited specialist access and high patient volumes, making accurate diagnostic support tools essential. Large language models like GPT-3 have demonstrated potential in clinical decision support but remain understudied in pediatric differential diagnosis.

Objective: This study aims to evaluate the diagnostic accuracy and reliability of a fine-tuned GPT-3 model compared to board-certified pediatricians in rural health care settings.

Methods: This multicenter retrospective cohort study analyzed 500 pediatric encounters (ages 0 - 18 years; n=261, 52.2% female) from rural health care organizations in Central Louisiana between January 2020 and December 2021. The GPT-3 model (DaVinci version) was fine-tuned using the OpenAI application programming interface and trained on 350 encounters, with 150 reserved for testing. Five board-certified pediatricians (mean experience: 12, SD 5.8 years) provided reference standard diagnoses. Model performance was assessed using accuracy, sensitivity, specificity, and subgroup analyses.

Results: The GPT-3 model achieved an accuracy of 87.3% (131/150 cases), sensitivity of 85% (95% CI 82% - 88%), and specificity of 90% (95% CI 87% - 93%), comparable to pediatricians' accuracy of 91.3% (137/150 cases; *P*=.47). Performance was consistent across age groups (0 - 5 years: 54/62, 87%; 6 - 12 years: 47/53, 89%; 13 - 18 years: 30/35, 86%) and common complaints (fever: 36/39, 92%; abdominal pain: 20/23, 87%). For rare diagnoses (n=20), accuracy was slightly lower (16/20, 80%) but comparable to pediatricians (17/20, 85%; *P*=.62).

Conclusions: This study demonstrates that a fine-tuned GPT-3 model can provide diagnostic support comparable to pediatricians, particularly for common presentations, in rural health care. Further validation in diverse populations is necessary before clinical implementation.

(JMIRx Med 2025;6:e65263) doi:10.2196/65263

KEYWORDS

natural language processing; NLP; machine learning; ML; artificial intelligence; language model; large language model; LLM; generative pretrained transformer; GPT; pediatrics

Introduction

The rapid advancement of artificial intelligence (AI) has led to the development of large language models (LLMs) that demonstrate sophisticated capabilities in understanding and analyzing human language [1]. Recent studies have shown promising applications of LLMs in health care, particularly in clinical decision support, medical knowledge synthesis, and diagnostic assistance [2-4]. However, their reliability and accuracy in specialized medical domains, especially pediatric care in resource-constrained settings, require thorough evaluation.

Differential diagnosis in pediatrics presents unique challenges that distinguish it from adult medicine. Young patients often cannot articulate their symptoms clearly, presentations can be atypical, and the range of potential diagnoses varies significantly with age. Recent systematic reviews have shown that diagnostic errors occur in "appreciable amounts" of pediatric encounters, with higher rates in rural and underserved areas [5]. These errors can lead to delayed treatment, inappropriate interventions, and potentially adverse outcomes.

The application of LLMs in clinical decision support has shown initial promise. Studies using GPT-3 and similar models have reported accuracies ranging from 75% to 85% in generating differential diagnoses for adult cases [6]. Notably, Steinberg et al [7] demonstrated that LLMs could achieve 82% accuracy in analyzing electronic health record (EHR) data for diagnostic support. However, pediatric applications remain underexplored, with limited studies specifically examining LLM performance in child and adolescent cases.

Rural health care settings face particular challenges that could benefit from LLM-based support tools. These areas often experience physician shortages, with providers managing high patient volumes and limited access to specialist consultation [8]. A survey of rural pediatric practices found that 52% of rural pediatricians report difficulty obtaining timely specialist input for complex cases [9]. Additionally, rural providers often work in isolation, managing a broad spectrum of conditions with fewer diagnostic resources compared to urban centers [10].

Previous evaluations of AI in pediatric diagnosis have largely focused on specific conditions or imaging-based applications rather than broad differential diagnosis. For instance, Wu et al [11] achieved 97.45% accuracy in pediatric otitis media interpretation using deep learning models, while other studies have demonstrated AI's effectiveness in detecting pediatric pneumonia from chest x-rays or identifying developmental disorders through automated screening tools. However, these models are often constrained by narrow diagnostic scopes, lack interpretability, and are not readily adaptable to general pediatric clinical reasoning.

Recent studies have begun to explore the application of LLMs in pediatric clinical settings. For example, Nian et al [12] found that ChatGPT and Google Gemini performed inadequately in providing recommendations for managing developmental dysplasia of the hip compared to expert guidelines, raising concerns about reliability in pediatric decision-making. Similarly, Wang et al [13] developed an LLM-based framework for pediatric obstructive sleep apnea management, highlighting the potential for specialized fine-tuning to improve diagnostic accuracy in specific pediatric conditions. Miyake et al [14] explored the role of AI-driven LLMs in pediatric surgery, emphasizing challenges related to real-time intraoperative decision support. Furthermore, Raza et al [15] investigated LLM applications in analyzing parental transcripts for children with congenital heart disease, demonstrating their potential role in augmenting thematic analysis in pediatric health care.

Despite these developments, comprehensive evaluations of LLMs in general pediatric differential diagnosis remain scarce. Many existing studies focus on narrow applications, lack real-world clinical validation, or fail to address age-specific nuances in pediatric presentations. Additionally, research on LLM utility in rural settings, where pediatricians may have limited access to specialist support, is particularly lacking. This study aims to bridge these gaps by systematically evaluating LLM performance in general pediatric differential diagnosis, with a focus on rural applicability and real-world clinical decision support.

The emergence of newer LLM architectures and their potential application in health care necessitates rigorous evaluation in real-world clinical settings [16]. While preliminary studies suggest promise, questions remain about their reliability, safety, and integration into clinical workflows [17]. Furthermore, the unique aspects of pediatric care—including age-specific disease presentations, developmental considerations, and the critical nature of early accurate diagnosis—require specific validation of these tools in pediatric populations [18].

This study addresses these knowledge gaps by evaluating the performance of a fine-tuned GPT-3 model in generating pediatric differential diagnoses within rural health care settings. By comparing the model's performance with that of experienced pediatricians across various age groups and presenting complaints, we aim to assess its potential as a clinical decision support tool. The findings could inform the development of AI-assisted diagnostic tools specifically tailored to the needs of rural pediatric health care providers.

Methods

Study Design and Setting

This multicenter retrospective cohort study was conducted in collaboration with a rural pediatric health care organization in Central Louisiana. The organization provides primary care to approximately 15,000 pediatric patients. The study analyzed patient data collected between January 2020 and December 2021. The overall workflow of the study is illustrated in Figure 1, encompassing data collection through model evaluation.

Figure 1. Workflow schematic showing the process of data collection, preprocessing, model training, and evaluation. The pipeline includes data splitting (70% training, 30% testing), GPT-3 fine-tuning, and comprehensive performance evaluation including subgroup analyses.

Ethical Considerations

Ethics approval was obtained from the Mansoor Pediatrics Ethics Committee (approval MP-2023‐017), and the study adhered to the principles of the Declaration of Helsinki. The study used retrospective, deidentified patient data and was exempt from informed consent requirements. Data were anonymized to ensure compliance with Health Insurance Portability and Accountability Act (HIPAA) regulations. No identifying information was accessible to researchers. No compensation was provided to participants as the study relied on existing retrospective data. For secondary analyses using deidentified data, the original consent obtained at the time of patient care covered the use of the data for research purposes.

Participants and Data Collection

A total of 500 pediatric patient encounters were included based on the following criteria:

- Inclusion criteria: Patients aged 0‐18 years with a documented chief complaint and pediatrician-generated differential diagnosis
- Exclusion criteria: Encounters with incomplete or inconsistent data

Anonymized data, including patient age, sex, chief complaint, presenting symptoms, medical history, and pediatrician-generated differential diagnoses, were extracted from the EHR system. Two independent researchers manually reviewed the data to ensure accuracy and consistency. No missing data were present in the final dataset. Demographic information, including racial and ethnic background, was not collected as part of this dataset. This omission limits the ability to assess potential biases in model performance across racial or ethnic groups, which is an important consideration for future research.

Five board-certified pediatricians (mean experience: 12, SD 5.8, range 5‐20 years) participated in the study as reference standard providers. Pediatricians were recruited from the participating health care organization based on their availability and experience in rural pediatrics.

Data Preprocessing

For each patient encounter, the chief complaint, presenting symptoms, and relevant medical history were concatenated into a single text string. Identifying information was removed to ensure privacy. Medical terms were standardized using a medical dictionary, and data were formatted for compatibility with the GPT-3 model.

Model Training and Fine-Tuning

The GPT-3 model (DaVinci version) was fine-tuned using the OpenAI application programming interface. The dataset was randomly split into a training set (n=350, 70%) and a testing set (n=150, 30%). The model was trained to generate up to five differential diagnoses for each input case. The study used retrospective data that included pediatrician-generated differential diagnoses documented during actual clinical encounters. No pediatricians were prospectively instructed to generate differential diagnoses specifically for this study. The same format of up to 5 differential diagnoses was used for standardization when processing both the historical physician documentation and the GPT-3 outputs. Fine-tuning parameters included 10 epochs, a batch size of 4, and a learning rate of 1e-5. The fine-tuning process aimed to optimize the model's ability to generate accurate and relevant differential diagnoses based on the input data. These details are visible in Multimedia Appendix 1.

GPT-3 (DaVinci version) was selected for this study because it was the most advanced version of the GPT model available at the time of data collection and model fine-tuning. Subsequent versions, such as GPT-3.5 and GPT-4, were released after the study period and were therefore not considered. Future work could explore the performance of these newer models in similar settings to assess potential improvements in diagnostic accuracy.

Evaluation Metrics

The model's performance was evaluated using the following metrics (Table 1):

- Accuracy: Proportion of correct predictions (true positives and true negatives) relative to total cases
- Sensitivity (recall): Proportion of actual positive diagnoses correctly identified by the model
- Specificity: Proportion of actual negative diagnoses correctly excluded by the model
- Precision: Proportion of positive predictions that were correct
- F_1-score: Harmonic mean of precision and sensitivity

In addition to these metrics, subgroup analyses were conducted by age group (0‐5, 6‐12, and 13‐18 years) and chief complaints (eg, fever, abdominal pain).

Table . Testing set evaluation metrics for analysis of the fine-tuned GPT-3 model, including formulas and values of the evaluation metrics for the GPT-3 model.

Metric	Formula	Description
Sensitivity (recall)	TP[a,b]/(TP + FN[c,d])	The proportion of actual positive diagnoses that were correctly identified by the model
Specificity	TN[e,f]/(TN + FP[g,h]) 0.90	The proportion of actual negative diagnoses that were correctly identified by the model
Precision	TP/(TP + FP)	The proportion of the model's positive predictions that were actual positive diagnoses
F_1-score	2 * (precision * sensitivity)/(precision + sensitivity)	The harmonic mean of precision and sensitivity, providing a balanced measure of the model's performance
Accuracy	(TP + TN)/(TP + TN + FP + FN)	The overall proportion of correct predictions made by the model

[a]TP: true positive.

[b]Cases where the model correctly predicted a positive diagnosis.

[c]FN: false negative.

[d]Cases where the model incorrectly predicted a negative diagnosis.

[e]TN: true negative.

[f]Cases where the model correctly predicted a negative diagnosis.

[g]FP: false positive.

[h]Cases where the model incorrectly predicted a positive diagnosis.

Statistical Analysis

Descriptive statistics were used to summarize patient demographics and model performance. χ^2 tests were used for categorical variables, and independent 2-tailed *t* tests were used for continuous variables. Statistical significance was set at $P<.05$. Data normality was assessed using the Kolmogorov-Smirnov test before statistical analysis. Our outcome metrics (accuracy, sensitivity, specificity) were found to follow a normal distribution ($P>.05$), supporting our use of parametric statistical methods including *t* tests for comparisons between groups. For nonnormally distributed variables, nonparametric alternatives (Mann-Whitney *U* test) were applied.

χ^2 tests were chosen for categorical variables due to their robustness in comparing proportions across groups. Independent *t* tests were selected for continuous variables after confirming normality of distribution. The choice of metrics (accuracy, sensitivity, specificity) aligns with standard diagnostic evaluation frameworks in health care AI validation studies. Subgroup analyses were performed to assess model performance consistency across demographics and clinical presentations, which is essential for evaluating potential biases in model predictions.

Power analysis indicated that a sample size of 500 would provide 80% power to detect a 10% difference in accuracy between the GPT-3 model and pediatricians, assuming a pediatrician accuracy of 90%. This calculation accounted for the expected distribution of common and rare diagnoses in our pediatric population, with consideration for potential subgroup analyses across different age groups and chief complaints.

Software and Tools

The statistical analysis was conducted using Python 3.8 (Python Software Foundation) [19] with the scikit-learn library [20] for model evaluation and SPSS Statistics version 29 (IBM Corp) for additional analysis [21]. The OpenAI application programming interface was used for model fine-tuning and prediction generation [22]. Software and scripts used in this study are available upon request for reproducibility.

Results

Dataset Characteristics

A total of 500 pediatric patient encounters were included, with 350 (70%) cases in the training set and 150 (30%) cases in the testing set. The mean age of patients was 7.5 (SD 5.2) years, and 52.2% (n=261) of participants were female. The most common chief complaints were fever (n=130, 26%), cough (n=98, 19.6%), abdominal pain (n=73, 14.6%), and rash (n=49, 9.8%). The distribution of age, sex, and chief complaint was similar between the training and testing sets (Table 2).

Table . Demographics and dataset characteristics.

Characteristic	Total (N=500)	Training set (n=350)	Testing set (n=150)	*P* value
Age (years), mean (SD)	7.5 (5.2)	7.4 (5.1)	7.7 (5.3)	.56[a]
Sex, n (%)				.82[b]
Female	261 (52.2)	184 (52.6)	77 (51.3)	
Male	239 (47.8)	166 (47.4)	73 (48.7)	
Chief complaint, n (%)				.93[b]
Fever	130 (26.0)	91 (26.0)	39 (26.0)	
Cough	98 (19.6)	70 (20.0)	28 (18.7)	
Abdominal pain	73 (14.6)	50 (14.3)	23 (15.3)	
Rash	49 (9.8)	34 (9.7)	15 (10.0)	
Other	150 (30.0)	105 (30.0)	45 (30.0)	
Rare diagnoses, n (%)	20 (4.0)	14 (4.0)	6 (4.0)	>.99

[a]*P* value calculated using independent 2-tailed *t* test.

[b]*P* value calculated using χ^2 test.

Model Performance

The fine-tuned GPT-3 model achieved high accuracy in generating differential diagnoses on the testing set. Key performance metrics are as follows:

- Accuracy: 87.3% (131/150 cases)
- Sensitivity (recall): 85% (95% CI 82% - 88%)
- Specificity: 90% (95% CI 87% - 93%)
- Precision: 89% (95% CI 86% - 92%)
- F_1-score: 0.87

The model correctly identified 128 positive diagnoses and excluded 334 negative diagnoses, with 16 false positives and 22 false negatives.

Subgroup Analysis

Performance across age groups and common chief complaints are summarized in Tables 2 and 3. The model's accuracy was consistent across age groups:

- 0 - 5 years: 87% (54/62 cases)
- 6 - 12 years: 89% (47/53 cases)
- 13 - 18 years: 86% (30/35 cases)

Table . Model performance by common chief complaints.

Chief complaint	Accuracy (95% CI)	Sensitivity (95% CI)	Specificity (95% CI)	Precision (95% CI)	F_1-score (95% CI)
Fever (n=39)	0.92 (0.88-0.96)	0.90 (0.85-0.95)	0.93 (0.90-0.96)	0.92 (0.87-0.97)	0.91 (0.86-0.96)
Cough (n=28)	0.89 (0.82-0.94)	0.85 (0.79-0.91)	0.90 (0.84-0.92)	0.89 (0.83-0.95)	0.87 (0.81-0.93)
Abdominal pain (n=23)	0.87 (0.78-0.92)	0.82 (0.75-0.89)	0.87 (0.83-0.90)	0.86 (0.79-0.93)	0.84 (0.77-0.91)
Rash (n=15)	0.93 (0.83-0.97)	0.88 (0.80-0.96)	0.91(0.88-0.94)	0.90 (0.92-0.98)	0.89 (81-0.97)

Similarly, the model demonstrated robust performance for common chief complaints:

- Fever: 92% (36/39 cases) accuracy
- Cough: 89% (25/28) accuracy
- Abdominal pain: 87% (20/23) accuracy
- Rash: 93% (14/15) accuracy

Subgroup analyses by age group and chief complaints revealed consistent performance, indicating the model's ability to generalize across varying pediatric presentations. However, the slight performance drop in complex and rare cases underscores the importance of targeted training datasets for improving diagnostic accuracy in these subgroups. For rare or complex diagnoses (n=20), the model achieved an accuracy of 80% (16/20 cases), slightly lower than the overall accuracy but comparable to pediatricians (17/20, 85% of cases; *P*=.62).

Comparison With Pediatricians

The model's performance was comparable to that of the 5 participating board-certified pediatricians. Pediatricians achieved an accuracy of 91.3% (137/150 cases), with a sensitivity of 92% (95% CI 91%-94%) and specificity of 88% (95% CI 84%-90%). Differences in sensitivity (*P*=.08) and specificity (*P*=.57) between the model and pediatricians were not statistically significant.

Statistical Analysis

χ^2 tests indicated no significant differences between the GPT-3 model and pediatricians for accuracy, sensitivity, or specificity. Subgroup analyses confirmed consistent performance across age groups and common chief complaints, with no significant performance disparities.

Tables

Table 1 provides a detailed breakdown of the evaluation metrics. Table 4 shows the performance of the model by age group, while Table 3 summarizes performance by chief complaints.

Table . Model performance by age group.

Age group (years)	Accuracy (95% CI)	Sensitivity (95% CI)	Specificity (95% CI)	Precision (95% CI)	F_1-score (95% CI)
Overall (n=150)	0.85 (0.81-0.89)	0.90 (0.87-0.93)	0.89 (0.86-0.92)	0.87 (0.83-0.91)	0.88 (0.85-0.91)
0 - 5 (n=62)	0.87 (0.82-0.92)	0.84 (0.79-0.89)	0.89 (0.85-0.93)	0.88 (0.83-0.93)	0.86 (0.81-0.91)
6 - 12 (n=53)	0.89 (0.84-0.94)	0.86 (0.81-0.91)	0.91 (0.87-0.95)	0.90 (0.85-0.95)	0.88 (0.83-0.93)
13 - 18 (n=35)	0.86 (0.80-0.92)	0.83 (0.77-0.89)	0.88 (0.83-0.93)	0.87 (0.81-0.93)	0.85 (0.79-0.91)

Discussion

Principal Findings

This study evaluated the diagnostic performance of a fine-tuned GPT-3 model in generating pediatric differential diagnoses in rural health care settings. The model achieved an accuracy of 87%, which was comparable to board-certified pediatricians' accuracy of 91%. Performance was consistent across age groups and common chief complaints, underscoring the model's potential as a reliable clinical decision support tool. While the model demonstrated lower accuracy for rare or complex cases (80%), its performance remained comparable to that of pediatricians (85%). These findings suggest that LLMs could enhance diagnostic accuracy and support providers in underserved regions, particularly for routine presentations.

Comparison to Prior Work

Our findings align with prior studies demonstrating the potential of LLMs in clinical decision support. For example, Steinberg et al [7] reported 82% accuracy in adult diagnostic support using LLMs, while Wu et al [11] achieved 97.45% accuracy in pediatric otitis media interpretation with deep learning models. This study extends these findings by focusing on general pediatric differential diagnosis, an area with limited prior research. Unlike previous studies that primarily examined urban or hospital-based datasets, our work highlights the utility of LLMs in resource-constrained rural environments, addressing a critical gap in the literature.

Strengths and Limitations

This study has several strengths. First, the use of real-world data from rural health care settings enhances the generalizability of findings to similar environments. Second, the inclusion of subgroup analyses provides insights into the model's performance across diverse age groups and chief complaints. Third, the comparative evaluation with experienced pediatricians underscores the model's clinical relevance.

Another of the key strengths of this study lies in its real-world applicability, particularly for rural health care settings where resources are limited and access to specialists is often constrained. By leveraging existing EHR data and evaluating the model's performance on common and rare pediatric conditions, this research provides a practical framework for integrating AI tools into primary care workflows. The consistent accuracy demonstrated across age groups and chief complaints highlights the potential of GPT-3 to serve as a valuable diagnostic support system for providers in underserved areas. However, implementing such tools in real-world clinical settings will require addressing infrastructure challenges, including internet connectivity and provider training. Despite these challenges, the findings underscore the feasibility of deploying AI systems to enhance diagnostic accuracy and reduce disparities in health care delivery, particularly in environments with high patient volumes and limited specialist availability.

However, there are notable limitations:

- Sample size and diversity: The sample size of 500 encounters, while informative, may not fully capture the diversity of the broader pediatric population. This limitation is particularly relevant in diverse health care settings, where factors such as demographic variability, socioeconomic status, and health care access can influence diagnostic patterns. Prior studies have demonstrated that models trained on limited datasets often fail to generalize across different populations, highlighting the need for larger, multi-institutional datasets to improve validity and applicability [17]. Additionally, our study used data from a single rural health care organization, which may limit the external validity of our findings. Similar studies have shown that AI-based diagnostic models exhibit performance degradation when applied to new patient populations due to variations in disease prevalence, clinical workflows, and physician documentation styles [18]. For instance, Steinberg et al [7] found that an LLM trained on one hospital's EHRs experienced a 15% drop in accuracy when tested on data from a different institution. These findings emphasize the need for external validation. Future research should prioritize expanding the sample size through multicenter collaborations, incorporating data from health care centers with diverse patient demographics to enhance generalizability and robustness. Similar initiatives have demonstrated improved AI model performance when trained on heterogeneous datasets, such as the multi-institutional validation study by Rajkomar et al [2], which improved diagnostic accuracy across multiple health care networks.
- Retrospective design: The use of retrospective data limits the ability to assess the model's impact on clinical workflows or patient outcomes. Prospective clinical trials are needed to evaluate these aspects.
- Cross-validation: A key limitation of this study is the lack of cross-validation across different health care organizations. Evidence suggests that AI-based diagnostic models frequently underperform when tested on external

datasets due to variations in clinical documentation, patient demographics, and institutional practices. For example, a systematic review of AI applications in health care found that models trained on single-center data exhibited an average 12% - 20% decrease in performance when applied to external datasets [17]. Steinberg et al [7] also demonstrated that LLMs trained on EHRs from one hospital struggled to maintain accuracy when exposed to unseen patient populations, emphasizing the importance of cross-validation. Furthermore, ChatGPT-based diagnostic models have shown variability in reliability across different patient demographics, particularly when applied to pediatric populations with rare conditions [12]. To ensure reproducibility, future studies should incorporate external validation using data from multiple institutions, including urban, suburban, and rural health care settings. By validating performance across diverse patient populations, we can assess the model's reliability in real-world clinical environments and mitigate the risks associated with dataset bias. This approach aligns with recommendations from previous research advocating for multicenter validation to improve AI model robustness [18].

- Rare diagnoses: The model's lower accuracy for rare or complex cases highlights the need for further fine-tuning and testing in these areas. Future fine-tuning efforts could incorporate domain-specific datasets, such as rare pediatric conditions or uncommon presentations, to enhance the model's diagnostic accuracy for less frequently encountered cases. For example, fine-tuning could focus on rare pediatric conditions such as Kawasaki disease or metabolic disorders, which often present atypically and are prone to diagnostic errors. Collaborations with specialist clinics could help build robust datasets for such conditions.
- GPT-3 versus newer models: Another limitation is the use of GPT-3 instead of its newer iterations, such as GPT-3.5 or GPT-4, which were released after the completion of this study. While GPT-3 demonstrated strong diagnostic performance, future studies should evaluate whether these more advanced models can further enhance accuracy, particularly for rare or complex cases. Specifically, GPT-3.5 and GPT-4 feature enhanced contextual understanding and larger training corpora [23], which may improve their ability to identify nuanced patterns in rare pediatric diagnoses. Additionally, these models may mitigate hallucination risks and offer better attribution of sources, which are critical for clinical applications. Comparative evaluations in similar rural health care settings would provide insights into their incremental benefits over GPT-3.

Practical Implications

Integrating LLMs like GPT-3 into rural health care settings could address critical challenges such as physician shortages, high patient volumes, and limited specialist access. These tools can provide rapid accurate diagnostic support, reducing diagnostic errors and improving patient outcomes [24]. However, practical barriers to implementation, including infrastructure requirements (eg, reliable internet and electricity) and provider training, must be addressed [25].

Reliance on AI systems poses risks, including overreliance by less experienced providers and challenges in managing incomplete or inconsistent input data [26]. Training programs should ensure health care providers understand the limitations of AI tools and develop strategies for validating AI-generated outputs. Establishing clear guidelines for AI use in clinical settings will further ensure patient safety and ethical application. To address concerns about hallucinations—instances where the model generates inaccurate or fabricated information—health care providers must verify AI-generated outputs against clinical guidelines and existing evidence. Integrating feedback mechanisms, where physicians can flag inaccuracies, may also help refine model behavior over time [27].

Additionally, fostering trust in AI tools among providers and patients will be essential for successful adoption [28]. Additionally, parental concerns regarding deferring diagnostic decisions to AI systems must be addressed to build trust and acceptance. Efforts to educate families about AI's role as a supplementary decision-making tool rather than a replacement for physician judgment are essential. Furthermore, rural health care facilities may face challenges in implementing AI solutions due to limited infrastructure, such as inconsistent internet access, power supply, and provider training [29]. These challenges may also include the cost of deploying and maintaining AI systems, as well as the need for ongoing technical support. Policy makers and health care administrators should explore subsidized programs or partnerships with technology providers to ensure equitable access to AI tools in resource-limited settings. Addressing these barriers will be crucial for ensuring successful adoption and integration into clinical workflows.

Future Directions

Future research should focus on the following:

- The findings should be validated in larger, more diverse populations across multiple health care settings.
- The diagnostic capabilities of more advanced models, such as GPT-3.5 or GPT-4, should be assessed to determine whether recent improvements in language model architecture further enhance diagnostic accuracy.
- The impact of LLM integration on patient outcomes, provider satisfaction, and workflow efficiency in prospective clinical trials should be assessed.
- User-friendly interfaces should be developed to facilitate adoption by providers with varying levels of technological expertise, and training programs tailored to rural health care providers should be developed to familiarize them with AI tools and address potential apprehensions about using such systems. These programs should emphasize the complementary nature of AI in clinical workflows rather than its replacement of human judgment.
- Ethical concerns, including data privacy, informed consent, and model transparency, should be addressed to ensure responsible use in clinical practice.
- In addition to traditional evaluation metrics, future studies should assess language generation issues such as hallucinations—instances where the model produces false or unsupported information—and attribution of responses to reliable sources.

These factors are critical for ensuring the safety and reliability of AI applications in clinical decision-making. Natural language processing metrics like Recall-Oriented Understudy for Gisting Evaluation (ROUGE) and bilingual evaluation understudy (BLEU) may be used to evaluate output quality, while further human review of generated responses could assess alignment with established clinical guidelines.

Conclusions

This study highlights the potential of GPT-3, a fine-tuned LLM, as a clinical decision support tool for pediatric differential diagnosis in rural health care settings. The model achieved diagnostic accuracy comparable to that of board-certified pediatricians, demonstrating robust performance across age groups and common presenting complaints. These findings suggest that LLMs could serve as valuable tools for addressing the unique challenges faced by rural health care providers, such as limited access to specialists and high patient volumes.

However, this work also underscores the need for further validation. Future research should focus on evaluating the model's performance in larger, diverse populations and real-world clinical settings. Ethical considerations, including data privacy and model transparency, must be prioritized to ensure responsible implementation. Another ethical consideration is the potential for AI models to exacerbate existing health disparities if their development does not account for diverse populations. Rigorous testing in underrepresented groups and ongoing audits for bias are critical steps to ensure fairness and equity in AI-driven health care applications. By addressing these challenges, LLMs like GPT-3 have the potential to enhance diagnostic accuracy, reduce disparities in access to care, and improve outcomes for pediatric patients in underserved regions.

While this study represents a step toward integrating AI into rural health care, its findings underscore the need for iterative improvements and cross-disciplinary collaboration to refine these tools. Partnerships between AI developers, clinicians, and health care administrators will be crucial in ensuring that AI solutions are both effective and accessible.

This study serves as a step in bridging the gap between AI innovation and practical health care applications, paving the way for future advancements in clinical decision support systems tailored to the needs of rural health care environments.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Technical appendix: GPT-3 Model specifications and implementation details.
[DOCX File, 24 KB - xmed_v6i1e65263_app1.docx]

References

1. Brown T, Mann B, Ryder N, et al. Language models are few-shot learners. In: Larochelle H, Ranzato M, Hadsell R, Balcan MF, Lin H, editors. Advances in Neural Information Processing Systems 33 (NeurIPS 2020): Neural Information Processing Systems Foundation, Inc; 2020:1877-1901.
2. Rajkomar A, Oren E, Chen K, et al. Scalable and accurate deep learning with electronic health records. NPJ Digit Med 2018 May 8;1:18. [doi: 10.1038/s41746-018-0029-1] [Medline: 31304302]
3. Liu Y, Ott M, Goyal N, et al. RoBERTa: a robustly optimized BERT pretraining approach. arXiv. Preprint posted online on Jul 26, 2019. [doi: 10.48550/ARXIV.1907.11692]
4. Cascella M, Montomoli J, Bellini V, Bignami E. Evaluating the feasibility of ChatGPT in healthcare: an analysis of multiple clinical and research scenarios. J Med Syst 2023 Mar 4;47(1):33. [doi: 10.1007/s10916-023-01925-4] [Medline: 36869927]
5. Rinke ML, Singh H, Heo M, et al. Diagnostic errors in primary care pediatrics: Project RedDE. Acad Pediatr 2018 Mar;18(2):220-227. [doi: 10.1016/j.acap.2017.08.005] [Medline: 28804050]
6. Hirosawa T, Kawamura R, Harada Y, et al. ChatGPT-generated differential diagnosis lists for complex case-derived clinical vignettes: diagnostic accuracy evaluation. JMIR Med Inform 2023 Oct 9;11(1):e48808. [doi: 10.2196/48808] [Medline: 37812468]
7. Steinberg E, Jung K, Fries JA, Corbin CK, Pfohl SR, Shah NH. Language models are an effective representation learning technique for electronic health record data. J Biomed Inform 2021 Jan;113:103637. [doi: 10.1016/j.jbi.2020.103637] [Medline: 33290879]
8. Marcin JP, Shaikh U, Steinhorn RH. Addressing health disparities in rural communities using telehealth. Pediatr Res 2016 Jan;79(1-2):169-176. [doi: 10.1038/pr.2015.192] [Medline: 26466080]
9. Pletcher BA, Rimsza ME, Cull WL, Shipman SA, Shugerman RP, O'Connor KG. Primary care pediatricians' satisfaction with subspecialty care, perceived supply, and barriers to care. J Pediatr 2010 Jun;156(6):1011-1015. [doi: 10.1016/j.jpeds.2009.12.032] [Medline: 20227727]
10. Chipp C, Dewane S, Brems C, Johnson ME, Warner TD, Roberts LW. "If only someone had told me…": lessons from rural providers. J Rural Health 2011;27(1):122-130. [doi: 10.1111/j.1748-0361.2010.00314.x] [Medline: 21204979]
11. Wu Z, Lin Z, Li L, et al. Deep learning for classification of pediatric otitis media. Laryngoscope 2021 Jul;131(7):E2344-E2351. [doi: 10.1002/lary.29302] [Medline: 33369754]

12. Nian PP, Umesh A, Jones RH, et al. ChatGPT and Google Gemini are clinically inadequate in providing recommendations on management of developmental dysplasia of the hip compared to American Academy of Orthopaedic Surgeons Clinical Practice Guidelines. J Pediatr Soc North Am 2025 Feb;10:100135. [doi: 10.1016/j.jposna.2024.100135]
13. Wang Q, Cao Z, Mao Q, et al. OSAer: a specialized LLM framework for pediatric obstructive sleep apnea management. SSRN. Preprint posted online on Jan 27, 2025. [doi: 10.2139/ssrn.5109772]
14. Miyake Y, Retrosi G, Keijzer R. Artificial intelligence and pediatric surgery: where are we? Pediatr Surg Int 2024 Dec 3;41(1):19. [doi: 10.1007/s00383-024-05921-8] [Medline: 39625492]
15. Raza MZ, Xu J, Lim T, et al. LLM-TA: an LLM-enhanced thematic analysis pipeline for transcripts from parents of children with congenital heart disease. arXiv. Preprint posted online on Feb 3, 2025. [doi: 10.48550/arXiv.2502.01620]
16. Bedi S, Liu Y, Orr-Ewing L, et al. Testing and evaluation of health care applications of large language models: a systematic review. JAMA 2025 Jan 28;333(4):319-328. [doi: 10.1001/jama.2024.21700] [Medline: 39405325]
17. Bedi S, Liu Y, Orr-Ewing L, et al. A systematic review of testing and evaluation of healthcare applications of large language models (LLMs). medRxiv. Preprint posted online on Apr 16, 2024. [doi: 10.1101/2024.04.15.24305869]
18. Rader B, Hswen Y, Brownstein JS. Further reflections on the use of large language models in pediatrics. JAMA Pediatr 2024 Jun 1;178(6):628-629. [doi: 10.1001/jamapediatrics.2024.0729] [Medline: 38683628]
19. Phillips D. Python 3 Object-Oriented Programming: Build Robust and Maintainable Software With Object-Oriented Design Patterns in Python 38: Packt Publishing Ltd; 2018.
20. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: machine learning in Python. J Machine Learning Res 2011 Oct;12:2825-2830 [FREE Full text]
21. Meyers LS, Gamst GC, Guarino AJ. Performing Data Analysis Using IBM SPSS: John Wiley & Sons; 2013.
22. Kublik S, Saboo S. GPT-3: The Ultimate Guide to Building NLP Products With OpenAI API: Packt Publishing; 2023.
23. Šimsová J. Examining cognitive abilities and multilingual performance of large language models: a comparative analysis of GPT-3 and GPT-4 [thesis]. : Charles University; 2024 URL: https://dspace.cuni.cz/handle/20.500.11956/195021 [accessed 2025-03-11]
24. Guo J, Li B. The application of medical artificial intelligence technology in rural areas of developing countries. Health Equity 2018 Aug 1;2(1):174-181. [doi: 10.1089/heq.2018.0037] [Medline: 30283865]
25. Khan F, Driessen J. Bridging the telemedicine infrastructure gap: implications for long-term care in rural America. Public Policy Aging Rep 2018 Nov 2;28(3):80-84. [doi: 10.1093/ppar/pry027]
26. Muley A, Muzumdar P, Kurian G, Basyal GP. Risk of AI in healthcare: a comprehensive literature review and study framework. Asian J Med Health 2023 Aug 28;21(10):276-291. [doi: 10.9734/ajmah/2023/v21i10903]
27. Reddy S, Allan S, Coghlan S, Cooper P. A governance model for the application of AI in health care. J Am Med Inform Assoc 2020 Mar 1;27(3):491-497. [doi: 10.1093/jamia/ocz192] [Medline: 31682262]
28. Sîrbu CL, Mercioni MA. Fostering trust in AI-driven healthcare: a brief review of ethical and practical considerations. Presented at: 2024 International Symposium on Electronics and Telecommunications (ISETC); Nov 7-8, 2024; Timisoara, Romania. [doi: 10.1109/ISETC63109.2024.10797264]
29. Olugboja A, Agbakwuru EM. Bridging healthcare disparities in rural areas of developing countries: leveraging artificial intelligence for equitable access. Presented at: 2024 International Conference on Artificial Intelligence, Computer, Data Sciences and Applications (ACDSA); Feb 1-2, 2024; Victoria, Seychelles. [doi: 10.1109/ACDSA59508.2024.10467443]

Abbreviations

AI: artificial intelligence
BLEU: bilingual evaluation understudy
EHR: electronic health record
HIPAA: Health Insurance Portability and Accountability Act
LLM: large language model
ROUGE: Recall-Oriented Understudy for Gisting Evaluation

Edited by A Schwartz; submitted 10.08.24; peer-reviewed by D Saderi, G Bender, T Olatoye, A Rahgozar, U Kumar Chalwadi, E Nwanaforo, P Hassan Ilegbusi, S Sakilay, M Collier; revised version received 24.02.25; accepted 28.02.25; published 19.03.25.

Please cite as:
Mansoor M, Ibrahim AF, Grindem D, Baig A
Large Language Models for Pediatric Differential Diagnoses in Rural Health Care: Multicenter Retrospective Cohort Study Comparing GPT-3 With Pediatrician Performance
JMIRx Med 2025;6:e65263
URL: https://xmed.jmir.org/2025/1/e65263
doi:10.2196/65263

Using Electrooculography and Electrodermal Activity During a Cold Pressor Test to Identify Physiological Biomarkers of State Anxiety: Feature-Based Algorithm Development and Validation Study

Jadelynn Dao[1*], BSc; Ruixiao Liu[2], BSc; Sarah Solomon[3], BSc, MD; Samuel Aaron Solomon[2*], BSc, MEng, PhD

[1]Computer Science, California Institute of Technology, Pasadena, CA, United States

[2]Medical Engineering, California Institute of Technology, 1200 E California Blvd, Pasadena, CA, United States

[3]Adult Psychiatry, Dartmouth College, Hanover, NH, United States

[*]these authors contributed equally

Corresponding Author:
Samuel Aaron Solomon, BSc, MEng, PhD
Medical Engineering, California Institute of Technology, 1200 E California Blvd, Pasadena, CA, United States

Related Articles:

Companion article: https://arxiv.org/abs/2411.17935v1

Companion article: https://med.jmirx.org/2025/1/e72093

Companion article: https://med.jmirx.org/2025/1/e77440

Abstract

Background: Anxiety has become a significant health concern affecting mental and physical well-being, with state anxiety (s-anxiety)—a transient emotional response—linked to adverse cardiovascular and long-term health outcomes. Traditional physiological monitoring lacks the contextual sensitivity needed to assess anxiety in real time. Electrooculography (EOG) and electrodermal activity (EDA), 2 biosignals measurable by wearables, offer promising avenues for identifying biomarkers of s-anxiety in naturalistic environments.

Objective: This study aims to identify novel biomarkers of s-anxiety using EOG and EDA signals collected in real-world conditions. We further explore how noninvasive wearable technology can enable real-time monitoring of physiological responses during induced stress, focusing on distinguishing true anxiety-related signals from artifacts in noisy environments.

Methods: Our study presents two datasets: (1) the EOG signal blink identification dataset Blink Identification Electrooculography Dataset (BLINKEO), containing both true blink events and motion artifacts, and (2) the EOG and EDA signals dataset Emotion, Electrooculography, and Electrodermal Activity Monitoring in Cold Pressor Conditions Dataset (EMOCOLD), capturing physiological responses from a cold pressor test (CPT). From analyzing blink rate variability, skin conductance peaks, and associated arousal metrics, we identified multiple new anxiety-specific biomarkers. Shapley additive explanations (SHAP) were used to interpret and refine our model, enabling a robust understanding of the biomarkers that correlate strongly with s-anxiety.

Results: BLINKEO feature analysis achieved a classification accuracy of 98.17% and F_1-score of 0.87 in distinguishing blinks from noise. In the EMOCOLD, survey results confirmed elevated anxiety and affectivity during the CPT, which normalized during recovery. SHAP analysis revealed that specific EDA features (eg, Hjorth activity and spectral entropy) and EOG features (eg, opening phase energy and signal height) consistently contributed to accurate predictions of s-anxiety and affectivity. Contextual combinations of features outperformed single-feature analyses, revealing relationships critical for robust biomarker identification.

Conclusions: These results suggest that a combined analysis of EOG and EDA data offers significant improvements in detecting real-time anxiety markers, underscoring the potential of wearables in personalized health monitoring and mental health intervention strategies. This work contributes to the development of context-sensitive models for anxiety assessment, promoting more effective applications of wearable technology in health care.

(JMIRx Med 2025;6:e69472) doi:10.2196/69472

KEYWORDS

stress; biomarker discovery; EOG; electrooculography; medical informatics; EDA; electrodermal activity

Introduction

Background

Despite being a short-term response, state anxiety (s-anxiety) has emerged as a significant factor impacting long-term health outcomes. Researchers have linked sustained s-anxiety with adverse cardiovascular effects [1], underscoring its profound effects on mental and physical health. Approximately 23.1% of US adults experience some form of diagnosable mental disorder [2], and 74% of US adults reported experiencing stress-related health issues within a given month [3], illustrating the widespread impact of anxiety-induced stress. Reliable biomarkers are essential for capturing the complexities of s-anxiety, enabling more precise and effective models.

Noninvasive wearable technology has the potential to transform health monitoring by continuously capturing physiological data through real-time sensor measurements [4,5]. These devices collect a broad array of metrics, yielding critical insights into the body's responses to anxiety. The ability to seamlessly collect large amounts of health-related data opens new ways to study and build an understanding of the onset and progression of anxiety, enabling more effective interventions and advancing our knowledge of human health. Identifying reliable biomarkers of s-anxiety offers a promising pathway to real-time health monitoring using wearable biosensors that can detect subtle physiological changes not immediately obvious in raw signal data.

The cold pressor test (CPT) is a widely used experimental method for studying anxiety responses in controlled settings. Participants immerse their hand in ice-cold water (0 - 4 °C), eliciting a sympathetic nervous system response. This test reliably induces physiological markers of anxiety [6-8], such as increased heart rate and sweat production. Other techniques, such as public speaking simulations and mental arithmetic tasks [9], also provoke anxiety and can be used to identify reliable biomarkers.

Physiological responses to s-anxiety and arousal have been extensively documented, revealing clear links between emotional states and indicators such as blink rate variability [10] and stress-induced sweating [11]. The 2-factor model of emotion, developed by Schachter and Singer [12], suggests that emotions arise from physiological arousal and subsequent cognitive interpretation. This model underscores that physiological responses are interpreted within a contextual framework, which are further hidden in indirect biomarkers for specific emotional experiences. For instance, fatigue, which affects the blink conditions, can intensify physiological arousal, directly impacting how the brain interprets anxious states. Such contextual cues are crucial for understanding s-anxiety in real-world settings, but they are often filtered out or controlled for in existing studies. Electrodermal activity (EDA) is a common measure of physiological arousal, but its reliability in depression research remains debated. Some studies report reduced EDA responses in individuals with major depressive disorder, suggesting impaired autonomic reactivity [13] and emotional hypo-responsiveness [14]. However, conflicting findings point to variability due to factors like medication use and methodological differences [15], emphasizing the need for further research on the relationship between physiological signals and emotional states.

Wearable devices offer a way to contextualize these arousal states dynamically. Through advanced human-machine interfaces, wearables can monitor how individuals respond to their environments, integrating data on physical responses to build a richer understanding of s-anxiety. There is growing interest in using noninvasive wearables to collect richer biomarker data for mental health study [16,17], interpreting physiological responses in respect to real-time contextual cues and providing a more comprehensive view of emotional states.

Research shows that blink rates tend to increase under difficult mental tasks or anxiety-provoking situations [18,19], reflecting activation of the autonomic nervous system. Electrooculography (EOG) captures electrical signals produced by eye movements, allowing for the detection of blink-related patterns. But EOG signals are often filtered out in stress studies to improve clarity of other signals [20], potentially overlooking valuable information related to emotional arousal. Studies suggest that specific components of EOG signals can be analyzed to extract physiological markers of s-anxiety, highlighting the need for further research into EOG biomarkers. Furthermore, fatigue—closely associated with emotional arousal—provides an additional avenue for understanding anxiety through EOG features [21,22]. Studies examining EOG signals in the context of drowsiness reveal correlations between blink frequency, blink duration, and stages of fatigue [19], highlighting a noninvasive method for tracking emotional arousal over time. Given the interplay between fatigue and anxiety, this relationship prompted our investigation into how fatigue-related features within EOG signals may serve as indirect indicators of anxiety, offering new opportunities for nuanced and comprehensive stress monitoring.

Similarly, stress has a pronounced effect on sweat production. Emotional sweating, triggered by the sympathetic nervous system, occurs in response to psychological stressors rather than temperature changes [15,16,23]. EDA is a method that measures changes in skin conductance. Under emotional arousal and stress, body sweats and skin conductance increases. Previous studies often rely on basic features like median values [24] or the phasic component of the EDA signal, focusing on nonspecific skin conductance responses (SCRs) to correlate with self-reported s-anxiety [25] scores. In such studies, peaks in the phasic signal exceeding 0.01 µS were counted as responses, and the frequency of these nonspecific SCRs per minute served as the primary measure for physiological s-anxiety. EDA primarily reflects the magnitude of emotional arousal without distinguishing between positive and negative affective states [26]. In other words, a high SCR could result from excitement or stress, making it challenging to interpret EDA data as a standalone indicator of anxiety. This underscores

the importance of using EDA in combination with other physiological markers [27], such as heart rate variability or blink rate, to gain a more comprehensive picture of an individual's emotional and physiological state. A more methodical exploration of signal characteristics found in EDA and EOG signals reveal nuanced physiological markers that strongly correlate with s-anxiety.

Currently, no widely accepted biomarkers reliably assess anxiety across diverse contexts, highlighting the need for continued exploration. Researchers have tested markers like heart rate variability, skin conductance, and blink rate, but results often vary due to individual differences and contextual influences. While many studies report that depressed patients exhibit reduced EDA responses, indicating diminished autonomic nervous system activity, some research presents conflicting findings. These discrepancies are attributed to variations in study designs, methodologies, and the influence of factors such as antidepressant treatment on EDA measurements [13].

While machine learning models have shown promise in detecting anxiety, their black-box nature limits interpretability, making it difficult to validate findings across diverse populations [28]. By introducing additional context-sensitive biomarkers, we aim to enhance the reliability and transparency of anxiety assessments, making models more applicable to real-world scenarios.

Objective

In our research, we leverage EOG and EDA data to develop a comprehensive, real-time model of s-anxiety. We have compiled 2 distinct datasets for this purpose. The first dataset, Blink Identification Electrooculography Dataset (BLINKEO), consists of EOG signal features from samples characterized by peak-like patterns, annotated to differentiate natural blink events from extraneous noise and wire movement artifacts. The second dataset, Emotion, Electrooculography, and Electrodermal Activity Monitoring in Cold Pressor Conditions Dataset (EMOCOLD), contains time-series EOG and EDA signals along with demographic data and stress responses elicited by the CPT. Using interpretability techniques such as SHAP (Shapley additive explanations), we identify and quantify specific biomarkers within the EOG and EDA data, with a focus on blink rate variability and sweat-related stress indicators. Our approach goes beyond simple anomaly detection by uncovering nuanced, anxiety-specific physiological markers informed by the 2-factor model of emotion. This research contributes to a more detailed understanding of stress mechanisms, with the potential to improve mental health interventions and enable personalized, context-specific stress management strategies with wearable technology.

Description of Question

This research aims to identify reliable, interpretable biomarkers of s-anxiety through EOG and EDA data for real-time stress monitoring.

Methods

Blink Identification EOG (BLINKEO) Data Collection

To create the BLINKEO dataset, EOG data were collected and analyzed to differentiate natural blinks from noise or wire movements. Our setup integrated the AD8232 (analog devices), a biopotential amplifier designed to capture physiological signals, which we optimized for measuring EOG activity. To detect vertical eye movements using EOG, one electrode was positioned above the eye and another below it, aligning on the vertical axis. This configuration captures the corneo-retinal potential changes associated with upward and downward eye movements. All trials were conducted on the same two individuals for consistency in signal characteristics. A total of 65 trials involving repeated blinking under controlled conditions where no extraneous movement occurred. In addition, 19 trials lasting between 30 seconds and 2 minutes were conducted under conditions with no blinking, but with deliberate wire movements introduced by manually adjusting or lightly tugging the electrode leads. These trials provided a baseline for accurately distinguishing noise artifacts from genuine blink events. Table 1 shows the characteristics of these trials, including session count, total recording time, and peak detection results before and after filtering.

To preprocess the EOG data, motion artifacts were identified and removed, to make the data suitable for downstream features. A fifth-order low-pass Butterworth filter using the Scipy Signal butter function was applied to isolate low-frequency components indicative of meaningful physiological signals. This was followed by a Savitzky-Golay filter using the Scipy Signal savgol_filter function for additional smoothing, which preserved essential features while reducing minor signal fluctuations [29].

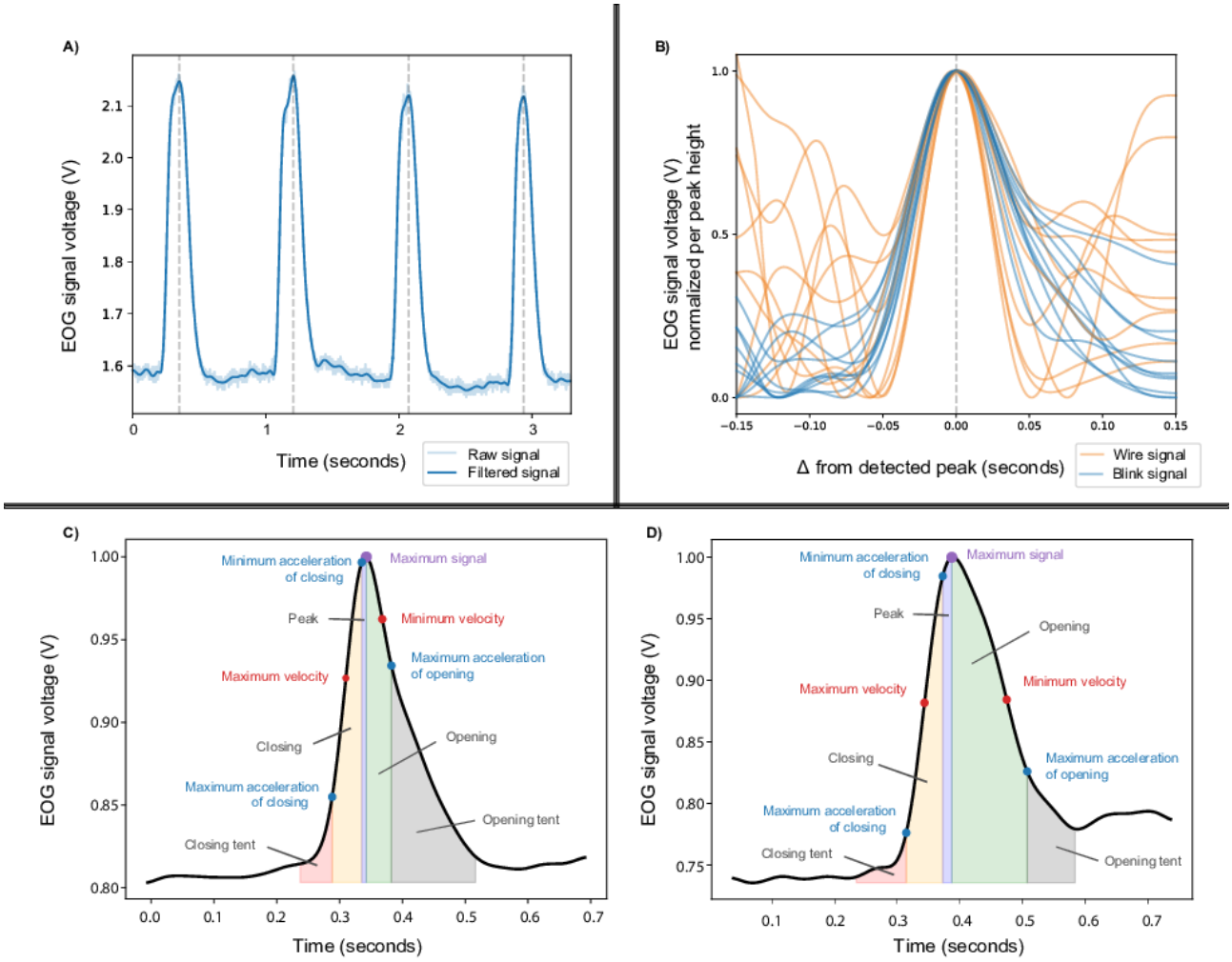
Table . Characteristics of blink and wire movement trials in the blink identification dataset. This table summarizes the number of independent sessions, cumulative recording time, and peak detection results before and after literature-supported blink peaks filtering for both blink and wire movement events.

Trial label	Sessions, n (%)	Total time (s)	Peaks detected, n (%)	Peaks after filtering, n (%)
Blink	65 (77)	12,103.14	6792 (54)	4734 (96)
Wire movement	19 (23)	2007.75	5704 (46)	203 (4)

Peak detection was performed using the Scipy Signal find_peaks function, identifying peaks with a prominence exceeding 0.1 with a peak width greater than 0.04 seconds [30] (blinks typically last between 0.1 and 0.4 seconds [31], averaging around 0.25 s). To focus on blink-like events, we additionally applied criteria based on established blink characteristics: a maximum peak width of 0.5 seconds and a minimum peak height of 0.05 volts [30]. We compared the signal quality after this initial peak detection with that obtained using conventional blink filtering methods. Traditional filtering techniques frequently overlook subtle blink patterns or introduce artifacts during data cleaning, potentially compromising accuracy. In

contrast, a learned-feature approach refines this process by reducing noise and enhancing the precision of true blink identification within the dataset. Figure 1A shows examples of detected blink peaks from the BLINKEO dataset, with red dotted lines marking the center of each peak. This figure demonstrates the effectiveness of the peak detection method described in this section, highlighting its ability to accurately locate and extract the central point of each blink event during blink trials.

Figure 1. A. Blink peak examples from the Blink Identification Electrooculography Dataset (BLINKEO). The grey dotted lines indicate the center of the peak, extracted by the peak detection method outlined in this section. B. Blink examples (blue) plotted against wire examples (green), as filtered EOG voltage signals, normalized per peak between 0 and 1. Peaks are time-aligned by time, in seconds, from the center of peak. Wire signals typically have higher variability. C. A singular blink peak. The purple dot marks the peak of the blink event, while the outer edges of the red and grey shaded sections represent the boundaries used for feature extraction. These boundaries are determined by identifying the nearest minimum on each side of the peak, providing a precise range for analyzing blink characteristics. D. Another example of a blink peak, demonstrating the variability in blink peak shapes observed across recordings. The feature extraction process remains consistent, with boundaries determined by identifying the nearest minima on either side of the peak. EOG: electrooculography.



However, wire movements can also produce peak-like shapes, which poses challenges for this filtering method. While effective in controlled or low-noise environments, the filter is easily triggered by noisy conditions, where artifacts such as wire movements may mimic blink patterns. Figure 1B presents time series segments of both blink and wire movement examples that have been classified as blinks under the current filtering approach, overlaid for comparison. The figure shows that wire movements exhibit greater variability in the regions surrounding the peak, as well as in the overall shape of the peak itself. Current approaches are unable to distinguish between true blinks and wire artifacts, underscoring the limitations of the method in noisier environments.

For each detected peak, baseline values were calculated to provide a reference point for the signal's amplitude. This involved locating the nearest minimum values on either side of the peak by performing binary search with a window size of up to 0.5 seconds in the left and right direction from the peak observed (see algorithm pseudocode in Multimedia Appendix 1). It recursively narrows down the search range to locate a local minimum, while avoiding minor fluctuations.

After establishing the baseline points, we extracted a comprehensive set of amplitude-independent features for each peak. These features include blink duration and various acceleration and velocity metrics, as used in previous EOG feature extraction and peak signal analysis studies [32,33]. A total of 32 peak-related features and label are stored as examples in the dataset, with labels distinguishing natural blinks from noise artifacts.

Figure 1C shows examples of EOG signals from 2 independent singular blink events, with distinct sections of the peak highlighted for clarity. The purple dot at the peak center represents the highest voltage point, detected by the peak detection algorithm. Red dots indicate local maxima in velocity, while blue dots show local acceleration points. Shaded regions in different colors represent key sections of the blink, such as the rising and falling phases, as well as acceleration and deceleration phases. This segmentation captures various aspects of the blink shape, this detailed segmentation provides valuable insights into the blink dynamics, enabling the extraction of relevant blink-related features.

We establish bounds for each feature by discretizing its range into 50 intervals. This discretization splits the feature's values into small, equally spaced segments, enabling a systematic exploration of possible lower and upper bounds that optimize model accuracy.

The process begins by identifying the minimum and maximum values of each feature. The range between these values is divided by the bin count (50), yielding an incremental "step size," or delta value, for testing. This delta value determines how much the threshold will shift at each iteration when exploring the bounds. To identify the best lower bound, the algorithm starts from the minimum value and iteratively adds the delta value (eg, 0.2) to the threshold, testing each increment by culling data points below it and evaluating the model's accuracy with the adjusted dataset. The lower bound with the highest accuracy is selected as the optimal starting point for that feature.

The search then proceeds to find an optimal upper bound, beginning with the maximum value and reducing it by increments of the delta value until reaching the previously identified lower bound. This decremental approach ensures the upper bound remains above the lower bound. Each new threshold is applied to the dataset, and the accuracy is recorded. The upper bound yielding the best accuracy becomes the final threshold for that feature.

The individually optimized lower and upper bounds for each feature are compiled into a list, representing the complete culling thresholds that maximize model performance across the dataset. By discretizing each feature's range into 50 intervals, the individual search method ensures a thorough yet efficient exploration of potential thresholds.

Emotion, EOG, and EDA Monitoring in Cold Pressor Conditions (EMOCOLD) Data Collection

The data collection process employed wearable sensors to record EDA and EOG signals from participants during controlled stress trials. EOG recording used the same setup as the BLINKEO data collection. Electrodes were positioned above and below one eye to detect vertical eye movements by capturing corneo-retinal potential shifts. EDA signals were recorded using a galvanic skin response sensor with MCP606 (microchip technology) operational amplifiers, operating at an excitation voltage of 0.5 V to measure skin conductance. Electrodes were placed on the forehead, chosen for its sensitivity to stress-induced sweat gland activity. The recorded signals were digitized and processed in real time using an ESP32-S3 WROOM-1 (Espressif Systems) microcontroller, which managed data acquisition, signal processing, and wireless transmission.

A total of 16 participants, between ages 26 and 31 years took part in the study, and demographic information, including race and sex, was collected and is summarized in Table 2. Data were taken from each subject only once. Each trial lasted about 10 - 15 minutes and was divided into 3 phases: baseline, CPT, and recovery. The length of the trial and the data used for feature analysis is as detailed in Table 3.

Table . Characteristics of trials in the Emotion, Electrooculography, and Electrodermal Activity Monitoring in Cold Pressor Conditions Dataset (EMOCOLD) dataset. Demographic details of the study participants, including race and assigned sex.

Characteristic	Count, n (%)
Assigned sex	
Male	11 (69)
Female	5 (31)
Race	
Asian	11 (69)
Hispanic or Latino	2 (13)
White	1 (6)
Middle Eastern or North African	1 (6)
Black or African American	1 (6)
Total participants	16 (100)

Table . Summary of trial durations across different experimental phases. Summary of the duration of time electrodermal activity and electrooculography features are collected across different experimental phases. For each phase—baseline (before hand submersion), cold pressor test (cold water immersion), and recovery (after hand removal)—the table lists the minimum, 25th percentile, median, 75th percentile, and maximum duration (in seconds).

Experiment	Length (seconds)		
	Minimum	Median (IQR)	Maximum
Trial			
Baseline	245.6	281.7 (274.0-310.0)	414.8
CPT[a]	261.9	290.4 (278.4-306.4)	358.0
Recovery	238.6	261.3 (252.8-278.1)	311.2
Feature collection			
Baseline	167.5	177.0 (172.1-182.3)	194.0
CPT	160.6	177.2 (165.0-184.1)	188.2
Recovery	157.1	172.1 (168.4-180.3)	191.9

[a]CPT: cold pressure test.

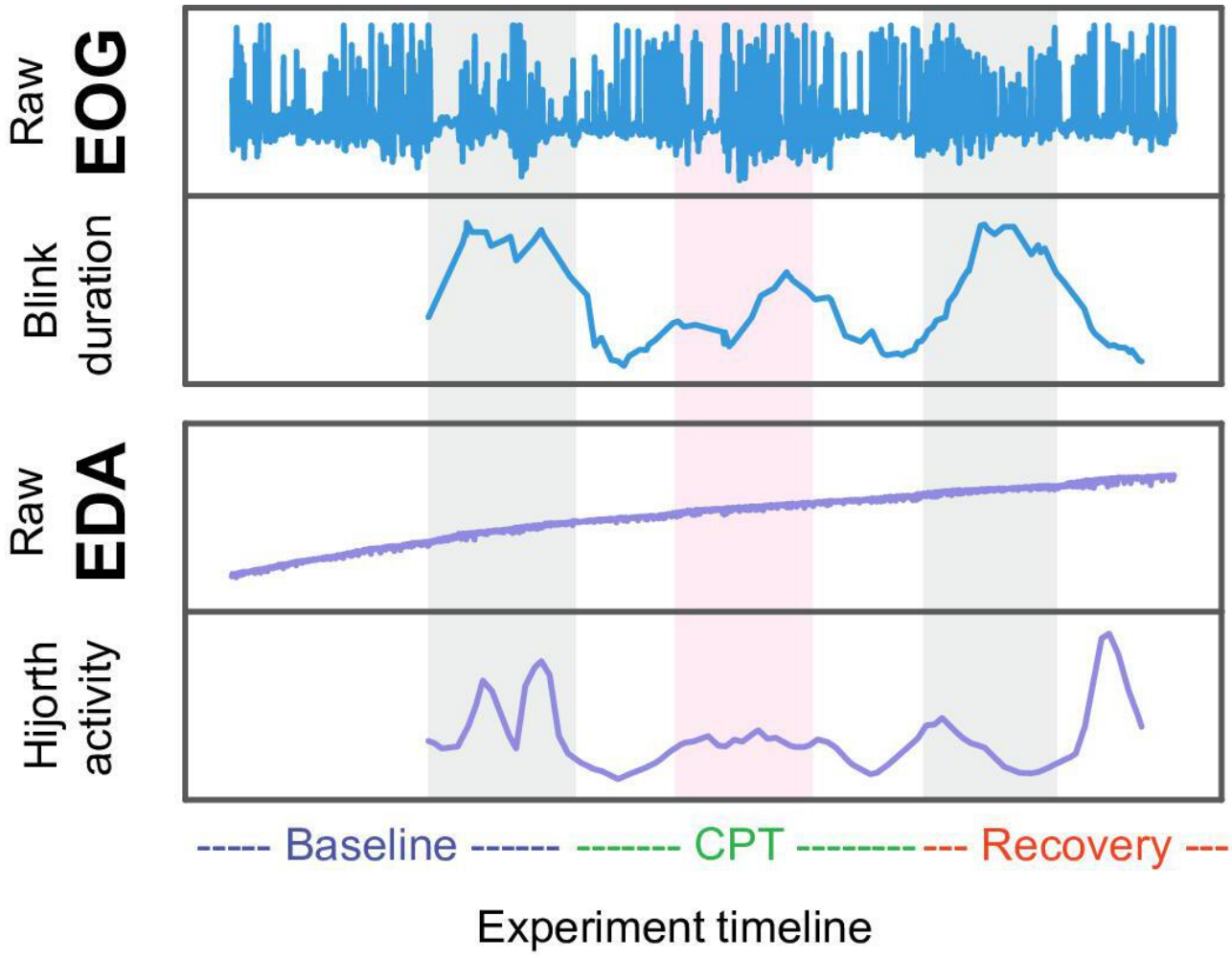
EOG signals were recorded using a 3-electrode configuration designed to capture vertical eye movements, particularly blink activity. Electrodes were positioned as follows: 1 above the eye, 1 below the eye, and a reference electrode in the middle of the forehead. This setup effectively captured vertical eye movement signals, with the reference electrode providing signal stability and reducing noise.

For EDA, a single electrode was placed on the forehead to measure changes in skin conductance associated with sympathetic nervous system activation. The forehead was chosen for its accessibility and stable conductance properties, making it suitable for detecting stress-related physiological changes in skin conductance.

Participants wore the device throughout the CPT trials, which were conducted to simulate acute stress events. The trials included both physical and environmental stressors. In the cold-water trials, participants immersed their hand in a circulating water bath set to a constant temperature of 0 - 6 °C. Participants maintained immersion for approximately 5 minutes or until voluntary withdrawal. This provided a controlled means of eliciting stress responses.

The design of these trials facilitated the collection of time-series data, capturing participants' physiological reactions to both physical exertion and environmental stressors, thereby providing a comprehensive view of their autonomic responses under varying stress conditions. Features were extracted from partitions of this sensor data, including statistical measures (mean [SD] and variance), signal entropy, peak detection metrics, and frequency-domain characteristics relevant to stress-induced physiological changes. Figure 2 shows a graphical depiction of the trial methodology.

Figure 2. This figure presents a visual representation of the experiment timeline and the signals recorded during the experiment, detailing the baseline, cold pressor test (CPT), and recovery phases. The raw electrooculography (EOG) and electrodermal activity (EDA) signals across these phases show no immediately clear trend distinguishing the baseline and recovery from the CPT stressor. However, when specific features such as blink duration from EOG and Hjorth activity from EDA are extracted and overlaid, more distinct patterns emerge, and can be used to quantify physiological responses to stress induction and subsequent recovery.



At each stage of the experiment—baseline, CPT, and recovery—participants completed an excerpt of the Positive and Negative Affect Schedule (PANAS) and the State-Trait Anxiety Inventory (STAI-State) to assess their emotional responses. The PANAS measures both positive emotions (eg, inspired and attentive) and negative emotions (eg, upset and nervous) on a 5-point scale, capturing general mood states. The STAI-State survey, consisting of items such as “I feel tense” and “I feel worried,” assesses immediate anxiety levels on a 4-point scale, making it particularly useful for tracking s-anxiety in response to acute stress. The survey recorded at each stage is detailed in Multimedia Appendix 2. Administering these surveys at each stage allowed us to correlate physiological data from EOG and EDA signals with subjective emotional responses, providing a comprehensive view of how participants’ mood and anxiety levels evolved across stress phases.

EOG Signal Segmentation

In analyzing EOG signals, we segmented the data to isolate individual blink peaks, which are essential for understanding blink dynamics in response to stress. From these peaks, we extracted 35 features, including blink duration, amplitude, frequency, and various acceleration and velocity metrics. A comprehensive list of these features and their definitions is provided in Multimedia Appendix 3.

EDA Signal Segmentation

The tonic and phasic components of skin conductance reveal different aspects of autonomic arousal, with the tonic level representing a stable baseline and the phasic response capturing transient, stimulus-driven changes. Tonic signals vary significantly across individuals due to factors like skin type and hydration, making them challenging to analyze consistently in relation to specific stress events. Phasic responses, however, reflect rapid fluctuations in skin conductance directly tied to acute stress or anxiety-inducing stimuli, characterized by quick rises and gradual declines.

Phasic signals were divided into rise and fall phases to capture the dynamics of the SCR, which is indicative of sympathetic nervous system activation. Specifically, peaks were detected by identifying rapid increases in skin conductance (rise phases) followed by gradual decreases (fall phases). To preprocess the EDA data and extract the phasic signal, motion artifacts were identified and removed, to make the data suitable for downstream features. A first-order low-pass Butterworth filter

was applied to isolate low-frequency components indicative of meaningful physiological signals.

This signal was divided into windows of 1 second in length. Each section was analyzed to determine key features, such as mean value, signal range, and standard deviation. 15 features were extracted from these windows, and the full list of features and their definitions can be found in Multimedia Appendix 4. These features are critical for quantifying the intensity and duration of autonomic arousal events, providing valuable insights into stress response dynamics. The segmentation process allowed for the extraction of detailed temporal characteristics of each skin conductance event, facilitating a comprehensive analysis of physiological arousal under stress.

Ethical Considerations

This study was conducted in accordance with ethical guidelines for research involving human participants. A total of 16 participants were recruited, following established ethical guidelines as delineated in protocols approved by the institutional review board at the California Institute of Technology (Caltech; protocol IR22-1280 and IR21-1102). Participants were not compensated. Participants were screened based on specific exclusion criteria, including non-English speakers unable to understand survey requirements, inability to provide informed consent, medication use affecting psychiatric states, pregnancy, irregular eye conditions (eg, ocular dysmetria), and pre-existing psychiatric or physical illnesses (eg, depression, anxiety, hypertension, hyperlipidemia, or chronic cardiovascular disease). All participants' data were fully anonymized, with identifying information removed and data transmission secured using byte-splicing encryption methods. The study adhered to data privacy and security protocols to ensure the confidentiality and protection of participants.

Results

Blink Identification EOG (BLINKEO) Analysis

Building upon the nonintentional blink signal processing outlined by previous research [34,35], a feature bounding analysis aligned closely with the study's approach of differentiating blink events based on slope and derivative features. By using blink duration alone as a feature, we achieved a classification accuracy of 87.46% and an F_1-score of 0.80 in distinguishing blinks from wire movements (see Multimedia Appendix 5). This suggests that feature extraction can yield strong performance metrics. Even without deep learning techniques, finding the right markers of blink peaks can reach the same efficacy of the study's outlined slope-based signal differentiation.

In our approach, we systematically evaluate all possible combinations of 5 selected features to optimize classification performance for distinguishing blink events from wire movements. For each feature combination, we apply a breadth-first search (BFS) traversal to explore and fine-tune the upper and lower bounds of each feature, seeking the configuration that maximizes classification accuracy.

The BFS traversal begins with initializing the bounds for each feature to cover its entire observed range, ensuring that no data points are culled at the outset. Each feature range is discretized into 15 bins, allowing for incremental adjustments to the bounds with a step size (delta) calculated as the range divided by the number of bins. These initial bounds are stored as a "node" in the BFS queue, representing a unique culling configuration.

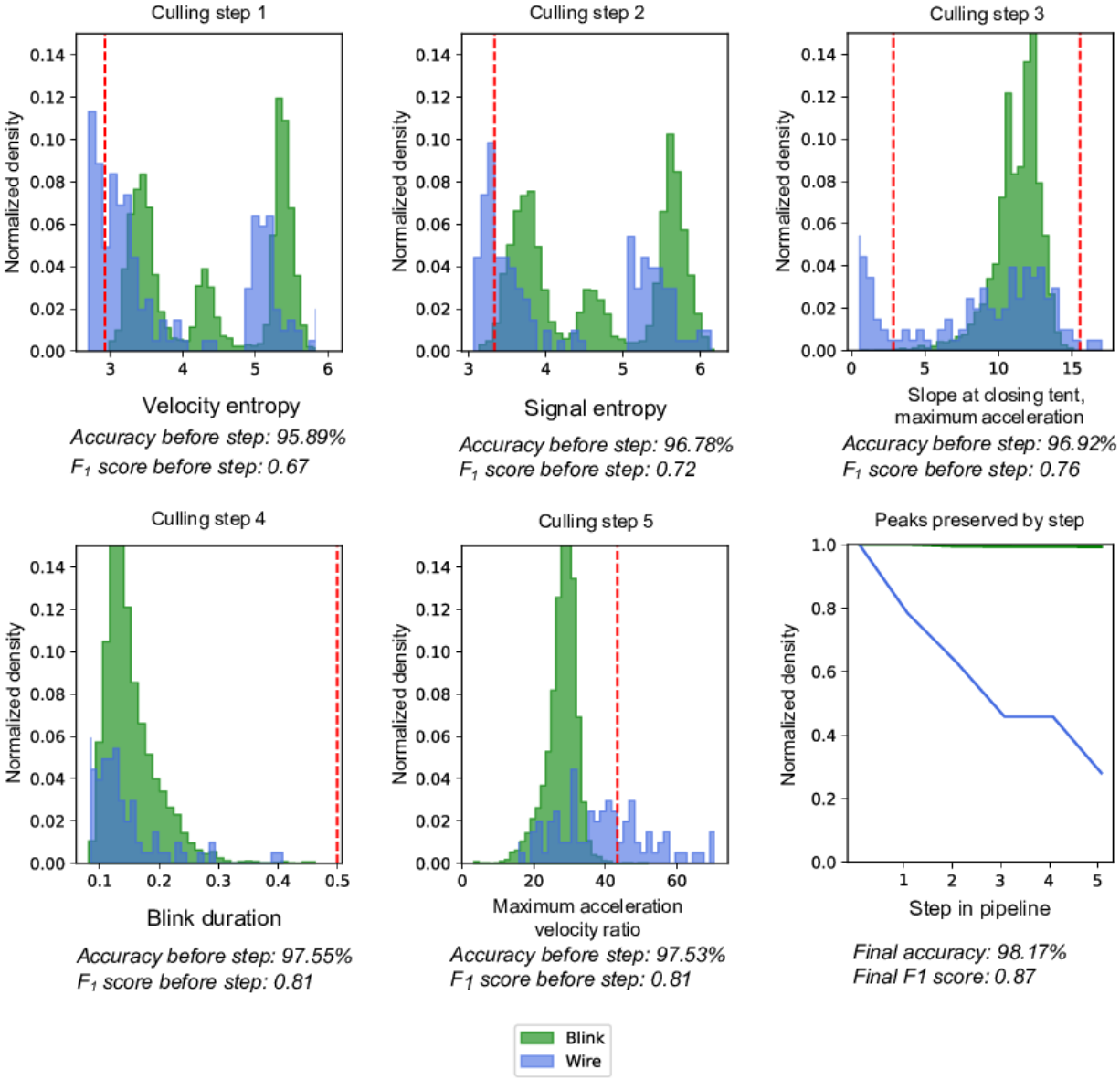
During each iteration of the BFS traversal, we dequeue a culling configuration and calculate its classification accuracy and F_1-score using a performance function. If the configuration achieves a higher accuracy than previously recorded, it becomes the current optimal configuration. The BFS traversal then generates neighboring configurations by slightly tightening the bounds for each feature—either increasing the lower bound or decreasing the upper bound by the computed delta. Each of these neighboring configurations, if unvisited, is added to the queue for further exploration.

This BFS traversal continues until all relevant bound configurations for the current feature combination are evaluated. The outcome is an empirically derived set of feature bounds that maximizes classification performance for each combination of features. By applying this process across all combinations of the selected 5 features, we ensure a comprehensive search of the parameter space, yielding an optimal culling pipeline tailored for precise blink detection. This method demonstrates the robustness of combining BFS with multifeature analysis to achieve a high-performing, data-driven classification model.

In our approach, we select combinations of 5 high-quality features and use a BFS traversal to optimize their combined bounds for maximal classification performance. For each combination, BFS systematically explores adjustments to the upper and lower bounds of each feature, identifying the optimal configuration that yields the highest accuracy and F_1-score.

The optimal feature combination achieved an accuracy of 98.17% and an F_1-score of 0.8734, using 5 key features that capture distinctive characteristics of blink dynamics. These features include velocity entropy, the entropy of the first derivative of the signal, which measures the variability and complexity of the blink motion; signal entropy, the entropy of the signal itself, providing a broader assessment of the overall blink pattern; slope at closing tent, maximum acceleration, the maximum acceleration during the closing phase of a blink, which isolates the rapid deceleration typical of blink closure; blink duration, representing the total time span of the blink event; and maximum acceleration velocity ratio, the ratio between the maximum acceleration and maximum velocity during the closing phase, which captures the relationship between these peak dynamics, indicative of voluntary eye closure. Figure 3 shows the results of each feature bounding step, against the BLINKEO labeled examples.

Figure 3. Optimal culling steps for differentiating blink events from wire movement artifacts in electrooculography (EOG) data. This figure presents the sequential culling steps optimized to achieve the highest accuracy and F_1-score in distinguishing blink events (green) from wire artifacts (blue) in EOG data. Each subplot demonstrates a unique culling step, applying specific feature thresholds to progressively refine the data. The final subplot, "Peaks preserved over culling pipeline," illustrates the proportion of retained peaks at each stage for both blink and wire signals, showcasing the efficacy of each step in isolating genuine blink events.



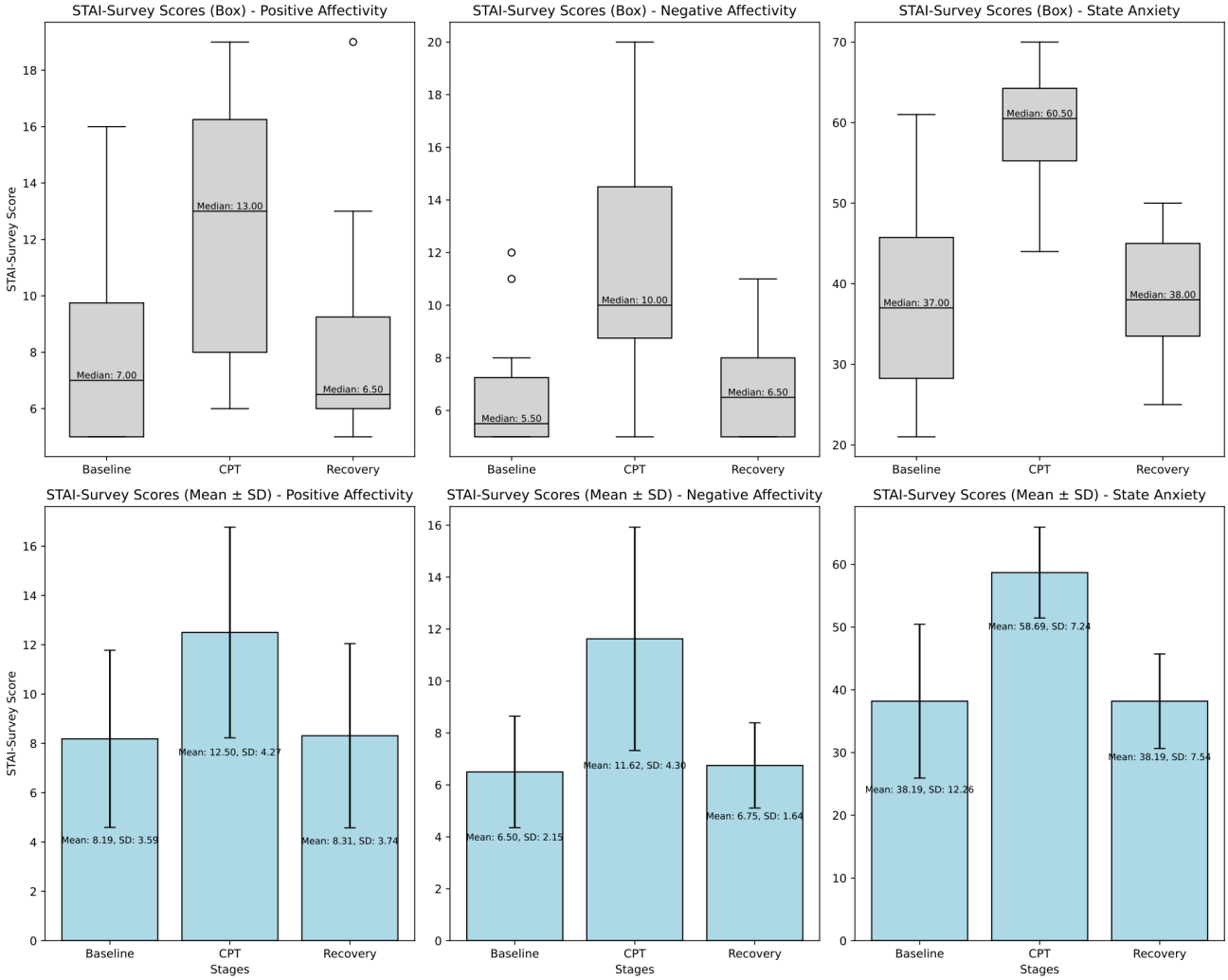
These features together form a comprehensive representation of blink characteristics, enabling differentiation of blinks from other signal types in the culling pipeline. This highlights how strategically selected bounds on multiple features, when combined, result in high classification performance without relying on complex algorithms.

Emotion, EOG, and EDA Monitoring in Cold Pressor Conditions (EMOCOLD) Analysis

Emotion Analysis

The EMOCOLD dataset analysis highlights significant physiological and emotional responses to acute stress induced by the CPT. Figure 4 shows participants' aggregated self-reported survey scores for positive affectivity, negative affectivity, and s-anxiety across the 3 trial stages: baseline, CPT, and recovery. Figure 4 shows that for each stage, survey responses were summarized and visualized using box plots, which display the distribution of scores. Positive affectivity and negative affectivity are scored on a scale of 5 - 25, and s-anxiety is scored on a scale of 20 - 80.

Figure 4. User-reported survey responses during each stage of the trial, displaying both box-and-whisker plots and column graphs for positive affectivity, negative affectivity, and state anxiety (s-anxiety) across the baseline, cold pressor test (CPT), and recovery stages. During the CPT, participants showed higher levels of positive affectivity, negative affectivity, and stage anxiety. Elevated levels recovered to baseline responses when participants took their hand out of the cold-water bath during the recovery phase. STAI: State-Trait Anxiety Inventory.



Participants reported increased positive and negative affectivity, as well as elevated s-anxiety during the CPT, which returned to baseline during recovery. This dual affective response suggests heightened arousal may include both alertness and discomfort. The recovery phase indicates effective autonomic regulation, as emotional states normalized once the stressor was removed. These findings validate the CPT as a method for inducing short-term anxiety.

SHAP Analysis

Overview

SHAP analysis is a method used to explain the output of machine learning models by breaking down the prediction into contributions from each feature. SHAP values are based on Shapley values from cooperative game theory, which attribute the impact of each feature on the model's output by treating each feature as a "player" in a game and calculating its contribution to the final prediction.

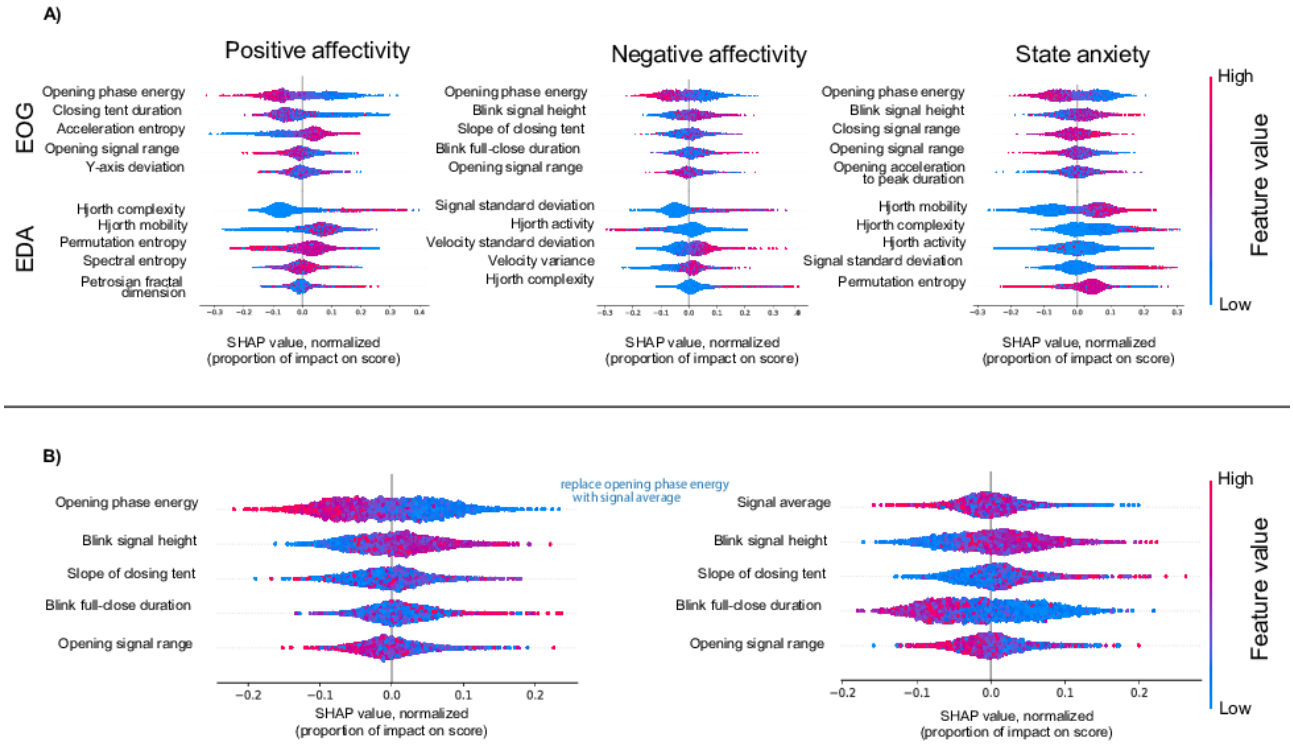
In this study, SHAP analysis was performed on combinations of 5 features, selected from the total feature set of 15 EDA and 35 EOG features, highlighting the significance of how certain biomarkers, used together, reveal more prominent interactions and effects on model predictions. This approach underscores that certain biomarkers, while potentially less impactful individually, can demonstrate substantial importance when analyzed as part of a group. By evaluating these interactions, we understand how combinations of features can provide insights into the model's behavior that single-feature analyses might overlook.

The quality of a set of features is determined by considering their collective contribution to the model's predictions, measured through the mean absolute SHAP values across the dataset. A high-quality set of features is one where the combination of features demonstrates substantial importance, as indicated by a higher mean absolute SHAP values. This benchmark reflects not only the magnitude of individual contributions but also the degree to which the features, as a group, interact to enhance the predictive power of the model.

The SHAP value maps provide insights into how various EOG features used in combination, and EDA features used in combination, contribute to predictions for positive affectivity negative affectivity, and s-anxiety. Each SHAP sub-plot illustrates the impact of individual features on model outputs, with higher SHAP values (toward the right) signifying a positive

contribution to the prediction, and lower SHAP values (toward the left) indicating a negative contribution. Figure 5A highlights the SHAP analysis identifying the combination of features that best polarize model predictions across the affective states.

Figure 5. 5A. Shapley additive explanations (SHAP) analyses for optimal combinations of 5 electrooculography (EOG) features (top row) and 5 electrodermal activity (EDA) features (bottom row) for positive affectivity (left column), negative affectivity (middle column), and state anxiety (right column). 5B. SHAP analysis of feature combinations. This analysis explores the quality of distinguishing different affectivity levels using different sets of features. This is an example of 5 EOG features and their impact on the negative affectivity score. Substituting one key feature with another can reveal new interdependencies among remaining features, thereby enhancing the model's interpretability.



EOG Feature Analysis

Among the EOG features analyzed, the opening phase energy, the integral of the opening phase of the peak signal, and opening signal range, the amplitude of the opening phase of the peak signal, consistently appeared in optimal feature combinations across all three outputs, suggesting their robustness as predictors. In addition, the signal height feature exhibited a particularly strong influence on predictions for negative affectivity and s-anxiety, underscoring its significance in these contexts.

EDA Feature Analysis

Among the EDA features analyzed, Hjorth parameters and the signal SD emerged as important predictors across the different affective states. These findings highlight the importance of analyzing feature interactions to reveal critical combinations that drive model performance, offering deeper insights into the physiological signals underpinning emotional and stress-related states.

The SHAP analyses in Figure 5B illustrate the importance of considering features in combination when identifying the most relevant biomarkers. By selecting sets of 5 features, we aim to identify a group of biomarkers that not only are individually relevant but also work effectively together. In Figure 5B, the inclusion of the feature opening phase energy contributes significantly to the model's performance, yielding a well-defined distinction in SHAP values. When opening phase energy is removed from the features considered, model performance decreases, and features such as blink full-close duration appear to show more distinction.

Discussion

Principal Findings

The main findings of this study show the potential of EOG and EDA as powerful tools for identifying nuanced physiological biomarkers associated with s-anxiety. Through the development and analysis of the BLINKEO and EMOCOLD datasets, we have introduced novel datasets and used advanced feature extraction techniques with interpretability methods such as SHAP analysis to uncover anxiety-specific markers. Our results emphasize the importance of understanding biomarkers in their context-dependent interactions and collective contributions to predictive models.

By systematically evaluating combinations of features, we mitigated challenges often faced in the literature, where biomarkers show inconsistent or nonsignificant correlations with anxiety due to situational variability. For instance, while blink rate and skin conductance metrics have been previously explored, our analysis reveals that their predictive use depends heavily on contextual factors, such as the type and intensity of the stressor. For example, biomarkers like blink duration and

skin conductance peaks performed well under controlled CPT conditions but may not generalize to other stress-inducing scenarios like public speaking. This underscores the need for adaptive, context-sensitive models that account for the situational variability of physiological responses.

A key contribution of this work is the identification of feature combinations that consistently provide reliable predictions. For EOG data, features like blink duration, peak height, and the opening integral were shown to be robust predictors across various emotional states. Similarly, for EDA data, features such as the mean signal, permutation entropy, and Hjorth activity emerged as significant contributors. By leveraging SHAP analysis, we identified not only which features are most relevant but also when and how they interact to enhance model performance. This approach offers a more comprehensive understanding of physiological responses compared to studies focusing solely on single-feature analyses.

Our findings bridge a critical gap in the literature by offering a systematic approach to addressing the variability and context-dependence of physiological biomarkers. This research advances the field by providing a framework for building more robust, interpretable, and context-sensitive models for anxiety assessment. The ability to dynamically adapt to different stress scenarios makes these biomarkers more applicable to real-world settings, paving the way for more personalized and effective mental health interventions.

Limitations

This study advances s-anxiety biomarker detection using EOG and EDA, but several limitations should be noted. The participant pool (N=16) was demographically skewed, with a predominance of male and Asian participants, limiting generalizability. Data were collected only once per subject, preventing analysis of intraindividual variability over time. Future studies should incorporate larger and more diverse populations with longitudinal data.

The CPT was conducted in a controlled lab environment, which may not fully reflect real-world anxiety triggers. In addition, motion artifacts in EOG recordings, despite filtering efforts, could impact signal clarity. EDA signals were recorded using a single forehead electrode, though different placements (eg, fingertips) may improve accuracy. Improved artifact detection and additional motion-tracking sensors could enhance data quality.

Feature selection for SHAP analysis focused on optimizing interpretability, but alternative selections may yield different insights. Models and analyses constructed using this dataset may not generalize well to other stress-inducing scenarios. External validation using independent datasets is necessary to confirm these findings.

Future Work

Future work should focus on validating these findings across diverse populations and stress-inducing contexts to further enhance the generalizability of these biomarkers. An important next step is to investigate potential gender-based and race-based differences in physiological responses to acute stress and our current methods of inducing stress, as this study was not explicitly designed for such analysis but acknowledges its relevance. In addition, integrating these models into wearable technology has the potential to revolutionize mental health monitoring, providing real time, personalized insights that could transform how we understand and manage anxiety. By addressing the challenges of situational variability and leveraging the strengths of combined biomarker analyses, this study contributes significantly to the growing field of wearable health technology and its applications in mental health.

Acknowledgments

This work was funded by the Translational Research Institute for Space Health through NASA NNX16AO69A, Office of Naval Research grants N00014-21-1-2483 and N00014-21-1-2845, Army Research Office grant W911NF-23-1-0041, National Institutes of Health grants R01HL155815 and R21DK13266, National Science Foundation grant 2145802, National Academy of Medicine Catalyst Award, and High Impact Pilot Research Award T31IP1666 from the Tobacco-Related Disease Research Program.

Generative artificial intelligence (AI; ChatGPT, version GPT-4o, OpenAI, 2025) was only used to assist with grammar correction and formatting of the authors' original text for this manuscript. These tools were used to improve clarity while preserving the authors' original ideas and content. All AI-assisted outputs were thoroughly reviewed and edited by the authors for accuracy and consistency before submission.

Data Availability

The datasets generated or analyzed during this study are available in the "stress-biomarkers-public-dataset" repository. The Blink Identification Electrooculography Dataset (BLINKEO) and Emotion, Electrooculography, and Electrodermal Activity Monitoring in Cold Pressor Conditions Dataset (EMOCOLD) can be accessed on GitHub [36,37].

Authors' Contributions

SAS and JD conceived the project. JD and SAS led the main study, collected the overall data, and contributed to the data analysis. RL and SS contributed to the platform development, characterization, human studies, and data processing. JD and SAS cowrote the paper. All authors provided feedback on the paper.

Conflicts of Interest

None declared.

Multimedia Appendix 1
The bounds of each peak were determined by performing 2 binary searches within the position domain of the signal—one to the left of the peak and one to the right.
[DOCX File, 123 KB - xmed_v6i1e69472_app1.docx]

Multimedia Appendix 2
The survey items from the Positive and Negative Affect Schedule (PANAS) and the State-Trait Anxiety Inventory (STAI-State) were used to assess participants' emotional and anxiety responses during the experiment. The PANAS scale consists of 10 items measuring positive affectivity and negative affectivity, each rated on a 1-5 Likert scale, where higher scores indicate stronger affective states. The STAI-State consists of 20 items assessing state anxiety, measured on a 1-4 Likert scale, where responses indicate varying degrees of agreement with statements reflecting anxiety levels. Higher scores in negative affectivity and anxiety-related items indicate greater distress, while higher scores in positive affectivity items indicate greater emotional well-being. The table below details each item, its corresponding scale, and the affectivity or anxiety dimension it evaluates.
[DOCX File, 16 KB - xmed_v6i1e69472_app2.docx]

Multimedia Appendix 3
Feature names and definitions extracted from windowed segments of electrodermal activity (EDA) signals.
[DOCX File, 15 KB - xmed_v6i1e69472_app3.docx]

Multimedia Appendix 4
Feature names and definitions extracted from windowed segments of electrooculography (EOG) signals.
[DOCX File, 17 KB - xmed_v6i1e69472_app4.docx]

Multimedia Appendix 5
A breadth-first search was performed to find the optimal range for distinguishing blinks from noise artifacts using the blink duration feature, which was extracted from electrooculography (EOG) signal peaks using our method. The identified bounds of 0.1227 to 0.3990 seconds align with values reported in the literature.
[DOCX File, 42 KB - xmed_v6i1e69472_app5.docx]

References

1. Steptoe A, Kivimäki M. Stress and cardiovascular disease. Nat Rev Cardiol 2012 Apr 3;9(6):360-370. [doi: 10.1038/nrcardio.2012.45] [Medline: 22473079]
2. Key substance use and mental health indicators in the United States: results from the 2022 National Survey on Drug Use and Health, center for behavioral health statistics and quality, HHS publication no PEP23-07-01-006, NSDUH series h-58. Substance Abuse and Mental Health Services Administration. 2023. URL: https://www.samhsa.gov/data/sites/default/files/reports/rpt42731/2022-nsduh-nnr.pdf [accessed 2024-11-03]
3. APA 2023 stress in America topline data. : American Psychological Association; 2023 URL: https://www.apa.org/news/press/releases/stress/2023/november-2023-topline-data.pdf [accessed 2024-11-12]
4. Kazanskiy NL, Khonina SN, Butt MA. A review on flexible wearables – recent developments in non-invasive continuous health monitoring. Sensors and Actuators A: Physical 2024 Feb;366:114993. [doi: 10.1016/j.sna.2023.114993]
5. Xu C, Song Y, Sempionatto JR, et al. A physicochemical-sensing electronic skin for stress response monitoring. Nat Electron 2024 Feb;7(2):168-179. [doi: 10.1038/s41928-023-01116-6] [Medline: 38433871]
6. Freeman R, Chapleau MW. Chapter 7 - testing the autonomic nervous system. In: Said G, Krarup C, editors. Handbook of Clinical Neurology, Peripheral Nerve Disorders: Elsevier; 2013, Vol. 115:115-136. [doi: 10.1016/B978-0-444-52902-2.00007-2]
7. Bullock T, MacLean MH, Santander T, et al. Habituation of the stress response multiplex to repeated cold pressor exposure. Front Physiol 2022;13:752900. [doi: 10.3389/fphys.2022.752900] [Medline: 36703933]
8. Mitchell LA, MacDonald RAR, Brodie EE. Temperature and the cold pressor test. J Pain 2004 May;5(4):233-237. [doi: 10.1016/j.jpain.2004.03.004] [Medline: 15162346]
9. Allen AP, Kennedy PJ, Dockray S, Cryan JF, Dinan TG, Clarke G. The trier social stress test: principles and practice. Neurobiol Stress 2017 Feb;6:113-126. [doi: 10.1016/j.ynstr.2016.11.001] [Medline: 28229114]
10. Bentivoglio AR, Bressman SB, Cassetta E, Carretta D, Tonali P, Albanese A. Analysis of blink rate patterns in normal subjects. Mov Disord 1997 Nov;12(6):1028-1034. [doi: 10.1002/mds.870120629] [Medline: 9399231]

11. Boucsein W. Principles of electrodermal phenomena. In: Electrodermal Activity: Springer US; 2012:1-86. [doi: 10.1007/978-1-4614-1126-0_1]
12. Schachter S, Singer JE. Cognitive, social, and physiological determinants of emotional state. Psychol Rev 1962 Sep;69(5):379-399. [doi: 10.1037/h0046234] [Medline: 14497895]
13. Sarchiapone M, Gramaglia C, Iosue M, et al. The association between electrodermal activity (EDA), depression and suicidal behaviour: a systematic review and narrative synthesis. BMC Psychiatry 2018 Jan 25;18(1):22. [doi: 10.1186/s12888-017-1551-4] [Medline: 29370787]
14. Dawson ME, Schell AM, Filion DL. The electrodermal system. In: Handbook of Psychophysiology, 3rd edition: Cambridge University Press; 2007:159-181. [doi: 10.1017/CBO9780511546396.007]
15. Carli V, Hadlaczky G, Petros NG, et al. A naturalistic, European multi-center clinical study of electrodermal reactivity and suicide risk among patients with depression. Front Psychiatry 2021;12:765128. [doi: 10.3389/fpsyt.2021.765128] [Medline: 35069276]
16. Castro Ribeiro T, García Pagès E, Ballester L, et al. Design of a remote multiparametric tool to assess mental well-being and distress in young people (mhealth methods in mental health research project): protocol for an observational study. JMIR Res Protoc 2024 Mar 29;13:e51298. [doi: 10.2196/51298] [Medline: 38551647]
17. Anmella G, Corponi F, Li BM, et al. Exploring digital biomarkers of illness activity in mood episodes: hypotheses generating and model development study. JMIR Mhealth Uhealth 2023 May 4;11:e45405. [doi: 10.2196/45405] [Medline: 36939345]
18. Korda AI, Giannakakis G, Ventouras E, et al. Recognition of blinks activity patterns during stress conditions using CNN and Markovian analysis. Signals 2021;2(1):55-71. [doi: 10.3390/signals2010006]
19. Thorslund B. Electrooculogram analysis and development of a system for defining stages of drowsiness [Master's thesis]. : Dept. Biomedical Engineering, Linköping University, LiU-IMT-EX-351; 2003 URL: https://www.diva-portal.org/smash/get/diva2:673960/FULLTEXT01.pdf [accessed 2025-11-12]
20. Mannan MMN, Kamran MA, Kang S, Jeong MY. Effect of EOG signal filtering on the removal of ocular artifacts and EEG‐based brain‐computer interface: a comprehensive study. Complexity 2018 Jan;2018(1). [doi: 10.1155/2018/4853741]
21. Zhu X, Zheng WL, Lu BL, Chen X, Chen S, Wang C. EOG-based drowsiness detection using convolutional neural networks. Presented at: 2014 International Joint Conference on Neural Networks (IJCNN); Jul 6-11, 2014; Beijing, China p. 128-134. [doi: 10.1109/IJCNN.2014.6889642]
22. Ma JX, Shi LC, Lu BL. An EOG-based vigilance estimation method applied for driver fatigue detection. Neurosci Biomed Eng 2014 Jun 1;2:41-51. [doi: 10.2174/22133852026661412181048 55]
23. Harker M. Psychological sweating: a systematic review focused on aetiology and cutaneous response. Skin Pharmacol Physiol 2013;26(2):92-100. [doi: 10.1159/000346930] [Medline: 23428634]
24. Sebastião R. Classification of anxiety based on EDA and HR. In: Goleva R, Garcia NDC, Pires IM, editors. IoT Technologies for HealthCare: Springer International Publishing; 2021:112-123. [doi: 10.1007/978-3-030-69963-5_8]
25. Strohmaier AR, Schiepe-Tiska A, Reiss KM. A comparison of self-reports and electrodermal activity as indicators of mathematics state anxiety. Frontline Learn Res 2020 Feb;8(1):16-32. [doi: 10.14786/flr.v8i1.427]
26. Horvers A, Tombeng N, Bosse T, Lazonder AW, Molenaar I. Detecting emotions through electrodermal activity in learning contexts: a systematic review. Sensors (Basel) 2021 Nov 26;21(23):7869. [doi: 10.3390/s21237869] [Medline: 34883870]
27. Gazi AH, Lis P, Mohseni A, et al. Respiratory markers significantly enhance anxiety detection using multimodal physiological sensing. Presented at: 2021 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI); Jul 27-30, 2021; Athens, Greece p. 1-4. [doi: 10.1109/BHI50953.2021.9508589]
28. Shatte ABR, Hutchinson DM, Teague SJ. Machine learning in mental health: a scoping review of methods and applications. Psychol Med 2019 Jul;49(9):1426-1448. [doi: 10.1017/S0033291719000151] [Medline: 30744717]
29. Virtanen P, Gommers R, Oliphant TE, et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat Methods 2020 Mar;17(3):261-272. [doi: 10.1038/s41592-019-0686-2] [Medline: 32015543]
30. Kleifges K, Bigdely-Shamlo N, Kerick SE, Robbins KA. BLINKER: automated extraction of ocular indices from EEG enabling large-scale analysis. Front Neurosci 2017;11:12. [doi: 10.3389/fnins.2017.00012] [Medline: 28217081]
31. Bartoshuk LM, Schiffman HR. Sensation and perception: an integrated approach. Am J Psychol 1977 Dec;90(4):718. [doi: 10.2307/1421748]
32. Reda R, Tantawi M, Shedeed H, Tolba MF. Analyzing electrooculography (EOG) for eye movement detection. In: Hassanien E, Azar AT, Gaber T, Bhatnagar R, Tolba MF, editors. The International Conference on Advanced Machine Learning Technologies and Applications (AMLTA2019): Springer International Publishing; 2020:179-189. [doi: 10.1007/978-3-030-14118-9_18]
33. López A, Ferrero F. Biomedical signal processing and artificial intelligence in EOG signals. In: Qaisar SM, Nisar H, Subasi A, editors. Advances in Non-Invasive Biomedical Signal Sensing and Processing with Machine Learning 2023:185-206. [doi: 10.1007/978-3-031-23239-8_8]
34. Zheng WL, Lu BL. A multimodal approach to estimating vigilance using EEG and forehead EOG. J Neural Eng 2017 Apr;14(2):026017. [doi: 10.1088/1741-2552/aa5a98] [Medline: 28102833]
35. Hassanein A, Mohamed A, Abdullah M. Classifying blinking and winking EOG signals using statistical analysis and LSTM algorithm. J Electr Syst Inf Technol 2023;10(1):44. [doi: 10.1186/s43067-023-00112-2]

36. Stress-biomarkers-public-dataset/blinkeo. GitHub. URL: https://github.com/jadee-dao/stress-biomarkers-public-dataset/tree/main/blinkeo [accessed 2025-06-04]
37. Stress-biomarkers-public-dataset/emocold. GitHub. URL: https://github.com/jadee-dao/stress-biomarkers-public-dataset/tree/main/emocold [accessed 2025-06-04]

Abbreviations

BFS: breadth-first search
BLINKEO: Blink Identification Electrooculography Dataset
CPT: cold pressor test
EDA: electrodermal activity
EMOCOLD: Emotion, Electrooculography, and Electrodermal Activity Monitoring in Cold Pressor Conditions Dataset
EOG: electrooculography
PANAS: Positive and Negative Affect Schedule
s-anxiety: state anxiety
SCR: skin conductance response
SHAP: Shapley additive explanations
STAI-State: State-Trait Anxiety Inventory

Edited by A Schwartz; submitted 30.11.24; peer-reviewed by D Saderi, S Rasania, T Olatoye, SM Savai, RSG Mahmoud, V Medeiros, M Collier; revised version received 10.05.25; accepted 12.05.25; published 10.07.25.

Please cite as:
Dao J, Liu R, Solomon S, Solomon SA
Using Electrooculography and Electrodermal Activity During a Cold Pressor Test to Identify Physiological Biomarkers of State Anxiety: Feature-Based Algorithm Development and Validation Study
JMIRx Med 2025;6:e69472
URL: https://xmed.jmir.org/2025/1/e69472
doi:10.2196/69472

Use of Mobile Forms in Low-Resource Areas for Population Health Surveys: Interview and Field Test Study

Alexander Davis[1,2*]; Aidan Chen[1,2*]; Milton Chen[2], PhD; James Davis[3], PhD

[1]University of California, Santa Cruz, CA, United States

[2]VSee Health, Newton, MA, United States

[3]Department of Computer Science and Engineering, University of California, 1156 High St, MS:SOE3, Santa Cruz, CA, United States

[*]these authors contributed equally

Corresponding Author:
James Davis, PhD
Department of Computer Science and Engineering, University of California, 1156 High St, MS:SOE3, Santa Cruz, CA, United States

Related Articles:

Companion article: https://arxiv.org/abs/2310.07888v1

Companion article: https://med.jmirx.org/2024/1/e64797

Companion article: https://med.jmirx.org/2025/1/e79539

Abstract

Background: Population health surveys are an important tool to effectively allocate limited resources in low-resource communities. In such an environment, surveys are often done by the local population with pen and paper. Data thus collected are difficult to tabulate and analyze.

Objective: The objective of this study was to evaluate the viability and efficiency of mobile forms as an alternative to paper-based surveys in a specific low-resource setting.

Methods: We conducted pilot interviews with 53 local surveyors in the Philippines to assess their initial attitudes toward mobile forms. We then built software that can generate mobile forms that are easy to use, capable of working offline, and able to track key metrics such as time to complete questions. Our mobile form was field-tested in 3 locations in the Philippines with 33 surveyors collecting health survey responses from 266 participants.

Results: In the pilot phase, we found that 32 out of 53 (60%) of the local surveyors preferred mobile forms over paper. After field-testing, the number of surveyors preferring mobile forms increased to 25 out of 33 (76%) after just using the form a few times. The mobile forms overall demonstrated enhanced efficiency in data collection and usability over paper surveys.

Conclusions: Our findings indicate that mobile forms are a viable method to conduct large-scale population health surveys in this low-resource environment.

(JMIRx Med 2025;6:e53715) doi:10.2196/53715

KEYWORDS

mobile forms; offline forms; electronic data capture; design; low-resource settings; health surveys

Introduction

There is an unmet need for medical care in low-resource communities, and telehealth clinics have the potential to reach patients in remote regions with insufficient coverage [1]. Our team and VSee interns have operated a series of free clinics with both in-person and telehealth physicians serving low-income populations in the Philippines with our partner organization Gawad Kalinga. Since Gawad Kalinga builds free housing in 10,000 locations across the Philippines, it can reach over 1 million households and mobilize many volunteers to support health initiatives, therefore letting us expand our ability to conduct large-scale health surveys [2]. Gawad Kalinga expressed interest in opening free clinics in all of its locations. We hypothesized that our mobile forms would be more efficient than traditional paper methods due to the ease of use of our software. Our goal was to evaluate whether mobile forms could

reduce time spent in data collection, enhance ease of use, and improve the accuracy of captured data [3].

In order to understand the target population and how best to serve them, we will perform a large set of surveys. Figure 1 provides an example of one community and our free clinic. The figure shows images of a street scene of the community, the interior of one of the health clinics, a remote doctor seeing a patient using telehealth, and a remote eye exam.

Figure 1. Images from one of our telehealth clinics in Manila, Philippines. The clinics were run by VSee, a US telehealth company. From left to right, the images portray the residential neighborhood in which the clinic took place, the interior of the clinic, a doctor (virtually on the laptop) seeing a patient, and a remote telehealth eye examination.

A good solution to conduct a large survey in a low-resource community can be challenging. Paper surveys are easy to use and record the data, but they would require a large amount of effort to transcribe information into digital format suitable for analysis [4]. Some health clinics adopt a parallel paper and digital format in an attempt to get the best of both [5]. Many organizations in affluent nations use dedicated tablets or laptops to conduct digital surveys, but in the Philippines, this solution is too expensive, and many people are not comfortable using such digital devices. A good platform solution to address these unique requirements is to use mobile phones already owned by the surveyors to conduct the survey [6].

The research aims to show that surveys conducted on mobile phones are feasible and efficient for data capture in low-resource and remote regions. We conducted pilot interviews to understand user requirements, built survey forms software, ran field-testing on surveys generated by the software, analyzed survey results, and performed postsurvey interviews.

Based on feedback from the pilot interviews, we determined that our mobile forms needed to satisfy 3 conditions. They should be extremely easy to use, allowing surveyors to complete the surveys with minimal training [7]. They should also function properly even if the internet went out. Finally, the survey should be able to measure the time required to type and complete questions. No existing mobile forms satisfied these conditions, so we created software that can generate this type of mobile forms.

A field study was performed to evaluate the practicality of our mobile forms. Over 3 days, in 3 separate locations, 33 surveyors, who were part of an initial group of 53, collected 20-question surveys from 266 respondents from the local population. The final count of 33 surveyors reflects those who were present for the postsurvey interviews, as 20 of the initial surveyors did not make it through the entire process.

After field-testing was complete, we interviewed the surveyors to evaluate user experience and satisfaction.

The contribution of our work is:

- Pilot interviews of surveyors
- Survey forms software: offline, easy, and timing aware
- Field-testing the mobile form–based surveys
- Analysis of field deployment

Methods

Pilot Interviews

We had little information on the background of our volunteer surveyors and how much knowledge they had in regards to using a phone, let alone completing a survey on one. A pilot interview to 53 surveyors was conducted to determine how comfortable the surveyors actually were with technology. We did this by manually interviewing each surveyor and asking them questions about preference and experience. Surveyors interviewed were chosen through convenience sampling based on their availability and proximity to survey sites. They were teens to seniors in their 70s, predominantly from low-income communities in Metro Manila.

Survey Forms Software

To evaluate mobile forms' performance and improve its design, we need to track information such as how much time surveyors take to complete each question. Therefore, a suitable mobile form must be easy to learn and use, able to work offline, and capable of tracking time spent.

While there are software solutions that can generate mobile forms for low-resource environments, none satisfied all of our needs. For instance, REDCap (Vanderbilt University) and SurveyMonkey (SurveyMonkey) can work offline and track response times, but their complexity or cost are not suitable for our use case, EpiCollect (Oxford University) [8] lacks the ability to track the time data we desired, Google Forms does not function offline, and ODK (University of Washington) [9] was deemed too complex by some stakeholders. We thus developed a software that generates mobile forms that can work offline with data analysis and visualization capabilities.

The mobile form software was developed using Next.js and React for the front end, with Redux for state management.

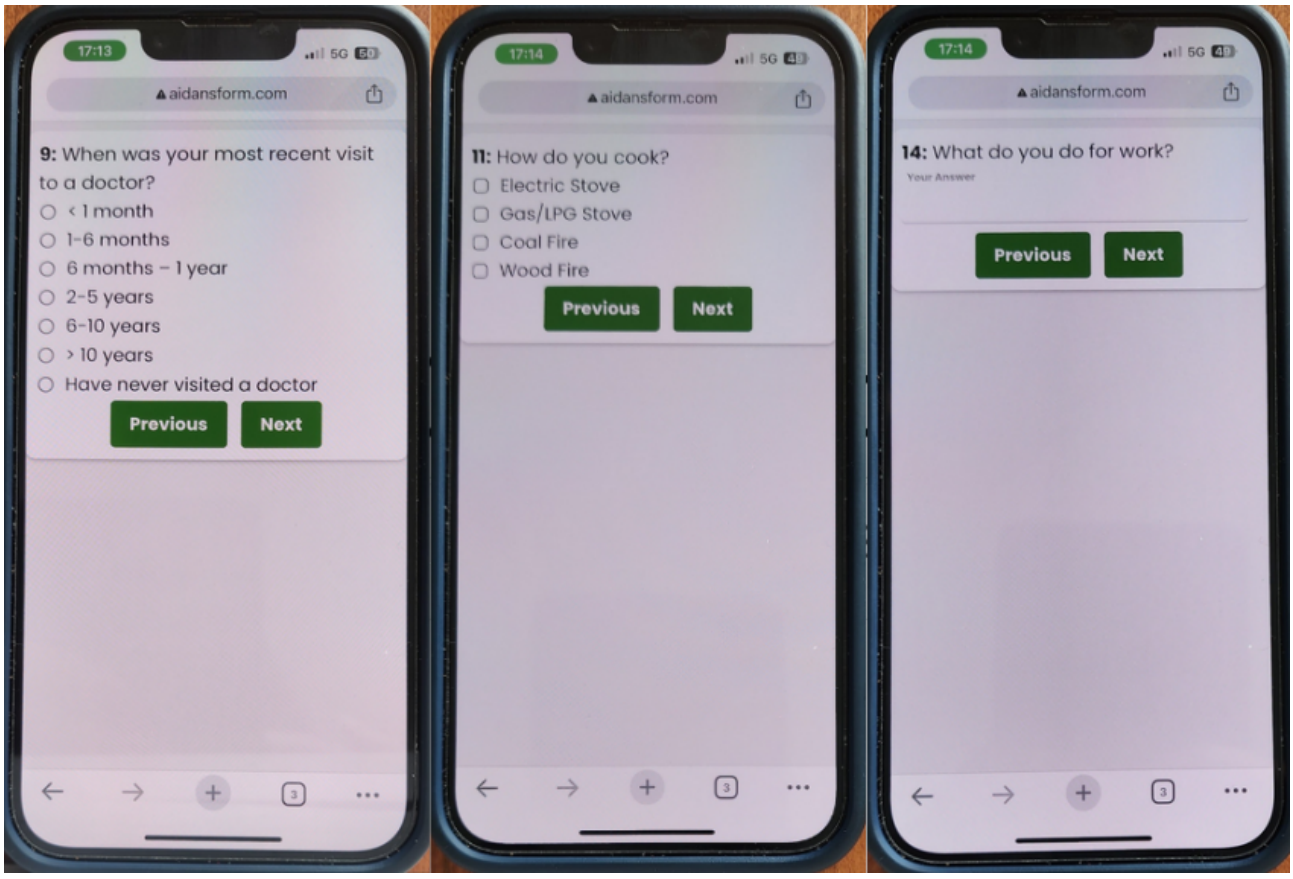
Firestore was used for real-time database functionality and Firebase (Google) handled back-end processes. Offline capabilities were managed through Redux's offline storage, ensuring data persistence even without internet access, and time-tracking functionality was implemented within the app's React components. The source code is available on GitHub under an open-source license to allow for replication and adaptation.

A feature of our mobile form is that it can work whether connected to the internet or not. It is common to experience unstable internet in low-resource environments where we collect data. At one of our previous clinics using Google Form, we observed intermittent internet outages to be sufficiently frustrating that some surveyors started writing on paper.

Another feature of our mobile form is that it tracks time data, so that we can study ease of use and quantify how each question will impact total effort required to collect a large number of surveys. The mobile form can measure time spent on each question, time spent typing the answer to each question, time the user spent to complete the survey, and the time that a user is not connected to the internet.

To encourage surveyors to use phones, we made our mobile form as easy and intuitive as possible. We followed design recommendations for low-resource populations such as minimization of visual complexity and streamlined navigation [10]. Our interface is shown in Figure 2. We presented only one question per screen to minimize scrolling and removed all extra interface buttons that might confuse users. Our goal was to make people feel comfortable using our mobile forms, and they would prefer it over paper.

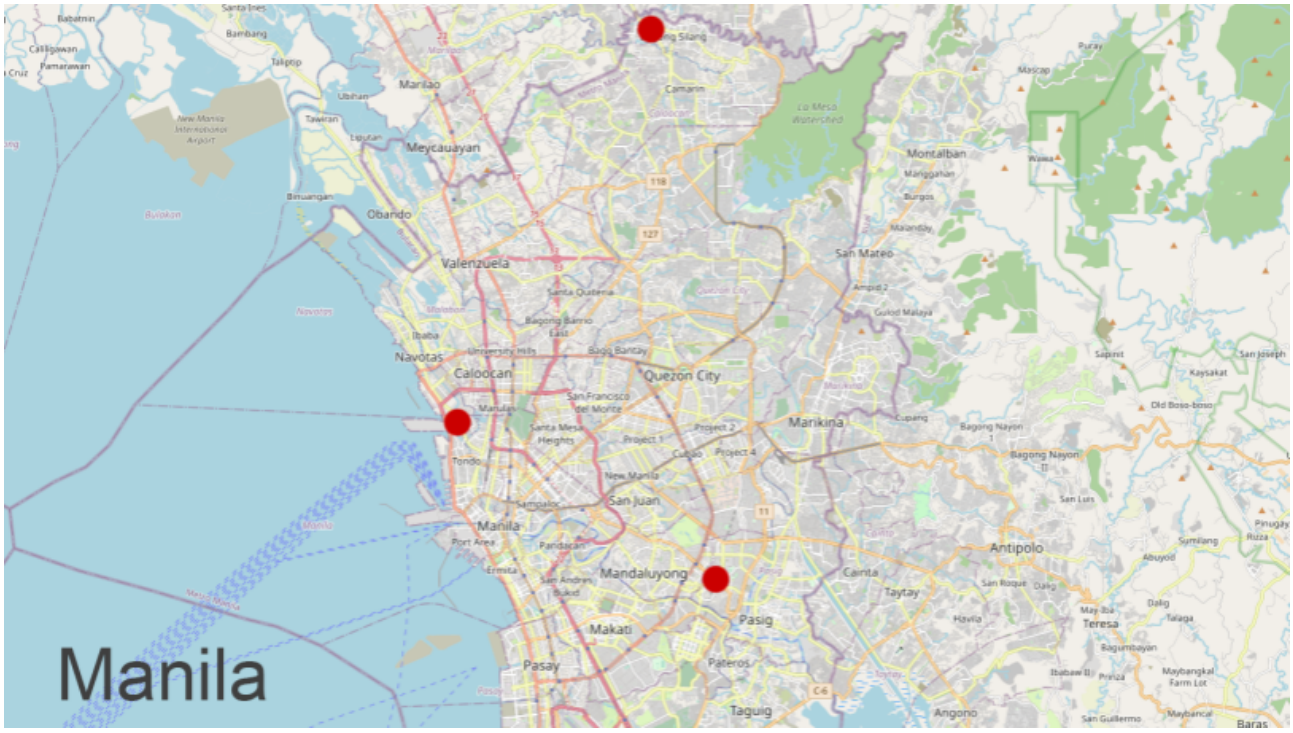
Figure 2. Mobile forms were created to work offline and track the time each question took the surveyors to complete. The interface was designed to be easy to learn and use, with only one question per loaded page.



Field Test

After designing a mobile form survey using our software, we tested its performance at 3 locations on 3 different days. The locations were in Metro Manila and ranged from an abject squatter settlement next to a trash dump to poor neighborhoods with access to water and electricity. Locations are shown on a map in Figure 3. The communities selected had between 600 and 1200 residents and were chosen to be representative of future survey sites.

Figure 3. Field tests occurred at 3 metro-Manila sites: 1 in Manila City and 2 in the neighboring communities of Caloocan and Pasig. A total of 33 volunteer surveyors participated, collecting 266 population health surveys.



We had 6 of our own staff training surveyors over 3 days. These surveyors were volunteers from local communities and selected to cover a wide age range. We were concerned that training might take a long time, but 96% of surveyors completed training in 10 minutes or less. The training covered basic phone usage and navigation of the mobile form.

The surveyors performed a total of 266 patient surveys over 3 days. The survey contained 20 questions related to health and access to health care. Some example questions are “What is the closest medical center or clinic that you would go to?” and “How do you get clean drinking water?” We wanted to find out which kind of questions were faster to complete for surveyors that had limited prior experience with mobile forms. The survey designed by our health team included both text and multiple choice questions.

Surveyors reported no issues with battery life or charging of power banks as the surveys were conducted over relatively short durations. This meant there was no need to provide additional power banks or batteries to ensure surveyors were able to complete their tasks effectively without interruptions.

Finally, we wanted to find out if the surveyors were satisfied with the mobile form they filled out. In order to accomplish this, we administered post–field test interviews to all 33 surveyors.

Ethical Considerations

This research paper reports retrospectively on prior interactions. The health clinic was run as part of the provision of social services, not specifically for this research project. The surveyors were volunteers with the clinic contracted for the purpose of administering surveys. All participants also had the option to withdraw at any time. Both of these activities were conducted in accordance with appropriate regulations, with all research procedures designed to align with the ethical principles outlined in the World Medical Association Declaration of Helsinki and adhered to relevant national and organizational standards for research involving human participants. The study reported in this paper made use of data that had already been collected in the course of normal operations, was anonymized, and finally provided to the researchers for analysis. Since this research was conducted on existing anonymized data, institutional review board approval and informed consent were not obtained specifically for this study.

Results

Pilot Interview Analysis

Our pilot survey asked several interview questions to better understand the target group. When asked what purpose surveyors used their phones for, surveyors indicated high use of social media (43/53, 81%) and messaging (39/53, 74%) apps on their phones, which may suggest a general level of comfort with mobile technology.

From our pilot interview, we learned that people had 2 concerns with mobile form surveys. First, a mobile form would be slower and more difficult to use than paper. Second, an unreliable network could render the form unusable. These 2 key concerns led us to creating our own custom survey forms software.

While we had anticipated the majority of surveyors preferring paper, only 40% (21/53) responded that they had a preference for paper, with the remaining 60% (32/53) preferring digital surveys. When asked why they preferred paper over digital surveys, the most common answer was that paper was “easier” or “faster” (17/21). Other answers included lack of phone ownership and concerns over poor internet access.

When observing how they typed, 32% (17/53) of surveyors typed with 1 finger, which may reflect a lack of experience and proficiency using modern digital phones. The rest typed with 2 fingers or used speed enhancements such as the autocomplete or swiping features on newer phones. The surveyors who typed with 1 finger preferred paper 53% of the time, while those using more advanced typing methods preferred paper 30% of the time. While we noticed a minor correlation between age and technical proficiency, as indicated by the typing method used, the data were insufficient to draw any definitive conclusions.

When asked if they had ever administered a population health survey before, more than 80% (43/53) responded no.

We concluded that many of the surveyors are not proficient modern phone users and have little experience in conducting health care surveys.

Field Test Analysis

In order to see how our surveyors actually performed when using the mobile form, we analyzed factors related to speed of completion of survey, which we hypothesized might be related to their reluctance to use mobile forms.

Table 1 shows the average time taken to complete each question. Notice that on average, multiple choice questions took a shorter amount of time to answer. Several of the questions could have been framed as either text or multiple choice, and based on this finding, we will encourage the health team to use multiple choice when possible.

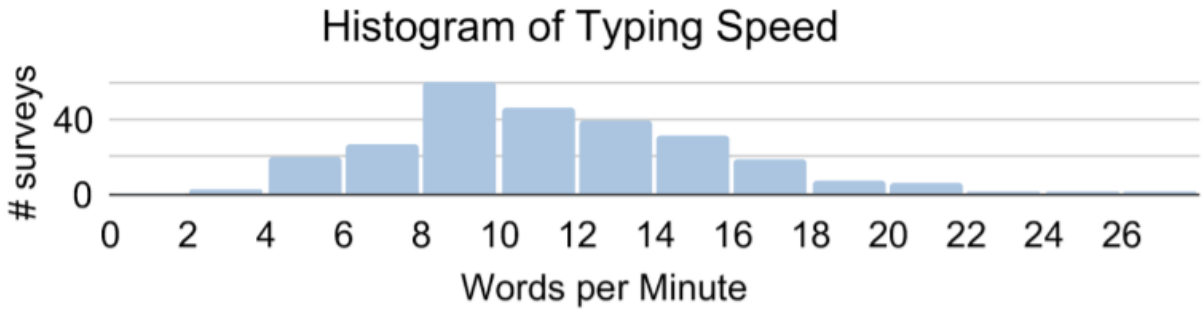
Average typing speed in textboxes across all 266 surveys was 11.5 words per minute; however, there was significant variation as shown in the histogram in Figure 4. This is consistent with previously reported typing speeds for low-resource populations [11].

Table . Average time to answer survey questions.[a]

Question	Type	Time (seconds)
Gender	Multiple choice	2.9
Type of phone owned?	Multiple choice	7.5
Source of clean drinking water?	Multiple choice	9.4
How do you cook?	Checkboxes	9.4
Lost wage for coming today?	Text	10.5
What do you do for work?	Text	11.9
How long does it take to go there?	Text	13.0
When did you last see a traditional healer?	Text	13.4
Cost to travel to a medical center?	Text	13.8
What is your education level, highest grade finished?	Text	14.5
About how much did it cost (roughly)? (Doctor's Visit)	Text	15.8
When was your most recent visit to a doctor?	Multiple choice	16.6
Date of Birth	Text	17.5
Your name (the volunteer doing the survey)	Text	18.5
What is the closest medical center or clinic that you would go to?	Text	20.4
Type of condition patient seeking evaluation and treatment	Checkboxes	21.9

[a]After field-testing was completed, we analyzed the time it took to complete each question. These data will be used to revise the survey to minimize the time needed to collect information.

Figure 4. Our mobile survey is capable of analyzing surveyor's typing speed. The average speed was 11.5 words per minute.



From the tracker that was built to record the time the internet was available while taking the survey, we found that 25% of the time the internet was unavailable during the survey. This indicates that a survey that can work offline is necessary.

Post–Field Test Interview

We collected data on whether surveyors found the mobile form easy to use. When asked "How easy was the mobile phone survey to use?", responses were very easy (n=23), somewhat easy (n=8), neither easy nor hard (n=2), somewhat hard (n=0), very hard (n=0). Of the 33 surveyors, 70% (n=23) agreed the mobile form was very easy to complete. No surveyors thought it was difficult to use. We also asked surveyors to describe positive attributes about mobile forms. A total of 30 out of 33 gave responses including the words "faster" or "easier," in contrast to opinions prior to using the forms. Results from these questions illustrate the effectiveness of our mobile forms.

Another question was the preference for paper or mobile forms. In the pilot interviews, we found that 60% (32/53) of the surveyors preferred mobile forms. In the post–field test interviews, we asked surveyors if they were to conduct another survey, would they use paper or mobile forms. 76% (25/33) responded that they preferred mobile over paper. Since the percentage of surveyors preferring mobile forms increased after the field test, we hypothesize that actual experience of using mobile forms for surveying improved their opinions. Since the surveyors on average only completed the mobile form 8 times during the field test, we believed with more usage, preference for mobile forms would continue to increase. The majority of surveyors expressed a strong preference for mobile forms, based on their user experience. While this feedback is subjective, it highlights the potential for mobile forms to be well received in future applications, though additional objective performance metrics would strengthen the validation of these findings.

Discussion

Principal Findings

Large-scale population health surveys are essential to deploy health care resources efficiently. We investigated the feasibility of using mobile forms to conduct such surveys in a low-resource environment in the Philippines. Initially, field team organizers requested to use paper to conduct the survey since many of the volunteer surveyors asked for it and there was the concern of an intermittent network. However, pilot interviews revealed a mobile form was actually preferred by 60% (32/53) of the surveyors. Based on insights gleaned from the pilot interviews, we built survey forms software that generated mobile form surveys. We then ran field trials to test the mobile form and conducted interviews afterward. The health survey was successfully completed using the mobile form. The percentage of surveyors preferring mobile forms increased to 76% (25/33) after just using the form a few times. The results demonstrate our mobile form is a viable method to conduct large-scale population health surveys in this low-resource environment.

Comparison to Prior Work

Several studies have already compared mobile-based surveys to paper-based data collection methods when it comes to health surveys. With an increase in the ownership of mobile phones, switching from paper surveys to digital surveys is becoming more appealing [12]. One study in Sudan found smartphone-based collection had fewer errors and faster retrieval than paper methods, seeing a reduction from an 83% error rate with paper questionnaires to a 17% error rate with smartphones [6]. In addition to that, data collected via smartphones were uploaded to the central database in a median time of 7 days, whereas paper-based data took a median of 21 days to be entered [13].

Our study builds on previous studies by developing a custom mobile form software, tested specifically in low-income environments in the Philippines. Our findings align with findings from a similar study done in rural Philippines using mobile apps [14]. Cost-effectiveness and reducing error rates have been focal points in recent studies, while implementing offline functionality has been studied less in recent studies. Our research advances this by developing custom mobile form software tailored for low-resource settings that is also extremely cost-efficient [15].

Unlike many studies that use existing platforms such as ODK, our custom mobile forms software offers a unique contribution by adding a focus on local surveyor needs for added equity when it comes to data collection. Our study further contributes to existing literature by providing insights into the implementation process of mobile forms in low-resource environments. The custom mobile forms software we developed addressed many of the common challenges identified in previous studies, such as the need for tools that can operate effectively in settings with limited or no internet connectivity [16].

Another example of an electronic data capture (EDC) framework designed for low-resource environments is ConnEDCt (Weill Cornell Medicine), a mobile EDC platform developed

specifically for clinical research applications in India and Ecuador. While their system also supported offline data collection and synchronization, their system was designed for more complex clinical research protocols including randomized controlled trials and regulatory-compliant data handling; however, our system focuses on a more simple and scalable approach for health surveys that can be set up quickly [17].

While there are a large range of tools in use, many do not offer a complete set of features and often require users to use multiple tools in parallel, thus complicating the workflows [7]. Our system fixes this issue as we consolidate every feature needed in a low-resource environment into one system in order to streamline the workflow.

Limitations

The study had several limitations that could influence the results of the study and its interpretation. First, the geographic scope of the study was restricted to Metro Manila, which limited the generalizability of the findings to other regions that could have different socioeconomic and technological conditions. Future studies could broaden the geographic scope to include more representation. Second, the sample size of the surveyors was small and future studies could use a larger and more diverse set of participants. To mitigate this limitation, we conducted pilot interviews to gather initial feedback, though future research should aim for larger scale studies to capture any potential variability. Third, the involvement of surveyors in testing the software may have introduced bias. While we tried to minimize biases by anonymizing survey responses and emphasizing the importance of honest feedback, future studies should consider using an independent group of testers. Fourth, technological familiarity among the surveyors, particularly the use of mobile phones, varied widely, potentially also affecting the usability testing. During our pilot interviews, we included questions that aimed to gauge the surveyors' comfort and proficiency with technology; however, we acknowledge that this initial assessment was not a comprehensive measurement of the surveyor's general literacy skills. Future studies should incorporate a more comprehensive evaluation of general literacy levels.

Data Availability

The datasets generated or analyzed during this study are not publicly available due to regulatory constraints associated with conducting research in a medical clinic, but are available from the corresponding author on reasonable request.

Authors' Contributions

AD, AC, MC, and JD contributed to the conceptualization. Data curation was conducted by AD and AC. AD, AC, MC, and JD performed the formal analysis. MC was responsible for funding acquisition. The investigation was carried out by AD, AC, MC, and JD. Methodology was developed by AD, AC, and MC. Project administration was handled by MC. Software was developed by AC. AD, AC, MC, and JD provided resources. Supervision was provided by MC and JD. Validation was performed by AD, AC, MC, and JD. Visualization was done by AD, AC, MC, and JD. The original draft was written by AD, AC, MC, and JD.

Conflicts of Interest

None declared.

References

1. Dorsey ER, Topol EJ. State of telehealth. N Engl J Med 2016 Jul 14;375(2):154-161. [doi: 10.1056/NEJMra1601705] [Medline: 27410924]
2. Pickering, MD A, Rafaqat, MD W, Balk A, et al. A building blocks approach to implementing a Telehealth Clinic Model to improve primary care access in the Philippines: a large-scale pilot project. Telehealth Med Today 2024;9(1). [doi: 10.30953/thmt.v9.456]
3. Mukasa O, Mushi HP, Maire N, Ross A, de Savigny D. Do surveys with paper and electronic devices differ in quality and cost? Experience from the Rufiji Health and demographic surveillance system in Tanzania. Glob Health Action 2017;10(1):1387984. [doi: 10.1080/16549716.2017.1387984] [Medline: 29157182]
4. Vaish R, Ishikawa ST, Liu J, Berkey SC, Strong P, Davis J. Digitization of health records in rural villages. Presented at: 2013 IEEE Global Humanitarian Technology Conference (GHTC); Oct 20-23, 2013; San Jose, CA. [doi: 10.1109/GHTC.2013.6713682]
5. Gainer A, Roth M, Strong P, Davis J. A standards-based open source application to gather health assessment data in developing countries. : IEEE Presented at: 2012 IEEE Global Humanitarian Technology Conference; Oct 21-24, 2012; Seattle, WA, USA p. 293-298. [doi: 10.1109/GHTC.2012.78]
6. Ahmed R, Robinson R, Elsony A, et al. A comparison of smartphone and paper data-collection tools in the Burden of Obstructive Lung Disease (BOLD) study in Gezira state, Sudan. PLoS ONE 2018;13(3):e0193917. [doi: 10.1371/journal.pone.0193917] [Medline: 29518132]
7. Silenou BC, Nyirenda JLZ, Zaghloul A, et al. Availability and suitability of digital health tools in Africa for pandemic control: scoping review and cluster analysis. JMIR Public Health Surveill 2021 Dec 23;7(12):e30106. [doi: 10.2196/30106] [Medline: 34941551]

8. Aanensen DM, Huntley DM, Feil EJ, al-Own F, Spratt BG. EpiCollect: linking smartphones to web applications for epidemiology, ecology and community data collection. PLoS ONE 2009 Sep 16;4(9):e6968. [doi: 10.1371/journal.pone.0006968] [Medline: 19756138]
9. Hartung C, Lerer A, Anokwa Y, Tseng C, Brunette W, Borriello G. Open data kit: tools to build information services for developing regions. Presented at: ICTD '10: Proceedings of the 4th ACM/IEEE International Conference on Information and Communication Technologies and Development; Dec 13-16, 2010; London, UK. [doi: 10.1145/2369220.2369236]
10. Khanna S, Ratan A, Davis J, Thies W. Evaluating and improving the usability of mechanical turk for low-income workers in india. Presented at: ACM DEV '10: First ACM Annual Symposium on Computing for Development; Dec 17-18, 2010; London, United Kingdom. [doi: 10.1145/1926180.1926195]
11. Gawade M, Vaish R, Waihumbu MN, Davis J. Exploring employment opportunities through microtasks via cybercafes. Presented at: 2012 IEEE Global Humanitarian Technology Conference; Oct 21-24, 2012; Seattle, WA. [doi: 10.1109/GHTC.2012.21]
12. Librero F, Ramos AJ, Ranga AI, Triñona J, Lambert D. Uses of the cell phone for education in the Philippines and Mongolia. Distance Educ 2007 Aug;28(2):231-244. [doi: 10.1080/01587910701439266]
13. Njuguna HN, Caselton DL, Arunga GO, et al. A comparison of smartphones to paper-based questionnaires for routine influenza sentinel surveillance, Kenya, 2011-2012. BMC Med Inform Decis Mak 2014 Dec 24;14(1):107. [doi: 10.1186/s12911-014-0107-5] [Medline: 25539745]
14. Kim TY, Baldrias L, Papageorgiou S, et al. A community-based survey to assess risk for one health challenges in rural Philippines using a mobile application. One Health Outlook 2022 Apr 5;4(1):7. [doi: 10.1186/s42522-022-00063-0] [Medline: 35379343]
15. Sundar DK, Garg S, Garg I, editors. Public Health in India: Technology, Governance and Service Delivery: Routledge; 2015.
16. Kenny A, Gordon N, Downey J, et al. Design and implementation of a mobile health electronic data capture platform that functions in fully-disconnected settings: a pilot study in rural Liberia. BMC Med Inform Decis Mak 2020 Feb 22;20(1):39. [doi: 10.1186/s12911-020-1059-6] [Medline: 32087731]
17. Ruth CJ, Huey SL, Krisher JT, et al. An electronic data capture framework (ConnEDCt) for global and public health research: design and implementation. J Med Internet Res 2020 Aug 13;22(8):e18580. [doi: 10.2196/18580] [Medline: 32788154]

Abbreviations

EDC: electronic data capture

Edited by T Leung; submitted 16.10.23; peer-reviewed by D Saderi, L Bert, Rakesh; revised version received 06.04.25; accepted 23.06.25; published 11.08.25.

Please cite as:
Davis A, Chen A, Chen M, Davis J
Use of Mobile Forms in Low-Resource Areas for Population Health Surveys: Interview and Field Test Study
JMIRx Med 2025;6:e53715
URL: https://xmed.jmir.org/2025/1/e53715
doi:10.2196/53715

Rapidly Benchmarking Large Language Models for Diagnosing Comorbid Patients: Comparative Study Leveraging the LLM-as-a-Judge Method

Peter Sarvari, MEng, MS, MBA; Zaid Al-fagih, BSc, MBBS, MPP

Rhazes AI, First Floor, 85 Great Portland Street, London, United Kingdom

Corresponding Author:
Peter Sarvari, MEng, MS, MBA
Rhazes AI, First Floor, 85 Great Portland Street, London, United Kingdom

Related Articles:

Companion article: https://www.preprints.org/manuscript/202409.0688/v3

Companion article: https://med.jmirx.org/2024/1/e69830

Companion article: https://med.jmirx.org/2025/1/e81235

Abstract

Background: On average, 1 in 10 patients die because of a diagnostic error, and medical errors represent the third largest cause of death in the United States. While large language models (LLMs) have been proposed to aid doctors in diagnoses, no research results have been published comparing the diagnostic abilities of many popular LLMs on a large, openly accessible real-patient cohort.

Objective: In this study, we set out to compare the diagnostic ability of 18 LLMs from Google, OpenAI, Meta, Mistral, Cohere, and Anthropic, using 3 prompts, 2 temperature settings, and 1000 randomly selected Medical Information Mart for Intensive Care-IV (MIMIC-IV) hospital admissions. We also explore improving the diagnostic hit rate of GPT-4o 05‐13 with retrieval-augmented generation (RAG) by utilizing reference ranges provided by the American Board of Internal Medicine.

Methods: We evaluated the diagnostic ability of 21 LLMs, using an LLM-as-a-judge approach (an automated, LLM-based evaluation) on MIMIC-IV patient records, which contain final diagnostic codes. For each case, a separate assessor LLM ("judge") compared the predictor LLM's diagnostic output to the true diagnoses from the patient record. The assessor determined whether each true diagnosis was inferable from the available data and, if so, whether it was correctly predicted ("hit") or not ("miss"). Diagnoses not inferable from the patient record were excluded from the hit rate analysis. The reported hit rate was defined as the number of hits divided by the total number of hits and misses. The statistical significance of the differences in model performance was assessed using a pooled *z*-test for proportions.

Results: Gemini 2.5 was the top performer with a hit rate of 97.4% (95% CI 97.0%‐97.8%) as assessed by GPT-4.1, significantly outperforming GPT-4.1, Claude-4 Opus, and Claude Sonnet. However, GPT-4.1 ranked the highest in a separate set of experiments evaluated by GPT-4 Turbo, which tended to be less conservative than GPT-4.1 in its assessments. Significant variation in diagnostic hit rates was observed across different prompts, while changes in temperature generally had little effect. Finally, RAG significantly improved the hit rate of GPT-4o 05‐13 by an average of 0.8% (P<.006).

Conclusions: While the results are promising, more diverse datasets and hospital pilots, as well as close collaborations with physicians, are needed to obtain a better understanding of the diagnostic abilities of these models.

(JMIRx Med 2025;6:e67661) doi:10.2196/67661

KEYWORDS

large language model; LLM; GPT-4; Gemini; Claude; retrieval-augmented generation; clinical medicine; diagnosis; diagnostic ability of LLMs; artificial intelligence; AI in medicine; AI in health care

Introduction

Background

In the United States alone, medical errors are the third largest cause of death [1], and within these errors, diagnostic errors result in the death or permanent disability of 800,000 people each year [2]. Research by The National Academy of Medicine [3] as well as Newman-Toker et al [4] estimated that diagnostic errors are responsible for approximately 10% of patient deaths [3,4] and 6% - 17% of hospital complications [3]. Moreover, 75% of diagnostic errors are cognitive errors [5], which are most commonly caused by premature closure and the failure to consider alternatives after an initial diagnosis has been established. Cognitive errors are also naturally linked to the overload and stress physicians experience, with current burnout rates reaching the highest ever levels recorded [6]. Given the recent progress in artificial intelligence (AI), large language models (LLMs) have been proposed to help with various aspects of clinical work, including diagnosis [7]. GPT-4, an LLM developed by OpenAI, has shown promise in medical applications with its ability to pass medical board exams in multiple countries and languages [8-11].

Comparing the Diagnostic Abilities of LLMs

Limited studies have attempted to compare the diagnostic abilities of LLMs and have mostly included (1) clinical vignettes; (2) case records directly from clinics; and (3) case reports, such as the *New England Journal of Medicine* (NEJM) Case Challenges. The latter are more complex than clinical vignettes and contain red herrings and other distractors to truly challenge a physician [12]. Khan and O'Sullivan [12] used 10 case challenges and compared diagnoses from GPT-3.5, GPT-4 (Bing), and Gemini 1.5 with the help of 10 physicians who filled out a grading rubric. The authors reported strong agreement among the graders who collectively preferred Gemini among the 3 models. Chiu et al [13] used 102 case records from the Massachusetts General Hospital and showed that GPT-4 outperformed Bard and Claude 2 in terms of diagnostic accuracy based on the *International Classification of Diseases, 10th Revision* (*ICD-10*) hierarchy. Shieh et al [14] asked GPT-3.5 and GPT-4 to analyze 109 USMLE (United States Medical Licensing Examination) Step 2 clinical knowledge practice questions (vignettes) as well as 63 case reports from various journals. The researchers concluded that while GPT-4 was 87.2% accurate on the vignettes, it was only able to create a shortlist of differential diagnoses for 47 of the case reports (75%). Other scholars have assessed the capabilities of various LLMs within a given specialty, such as otolaryngology [15] and radiology [16].

Many authors have focused on evaluating the diagnostic ability of a single LLM: GPT-4 was the most popular choice as it was generally the most accurate LLM at the time. Eriksen et al [17] asked GPT-4 to choose 1 of 6 diagnostic options for each of 38 NEJM case challenges, whereas Kanjee et al [18] relied on NEJM clinicopathological case conferences and tasked GPT-4 to first state the most likely diagnosis and then give a list of differentials. Manual review by the authors concluded that in 45 out of the 70 cases, the correct answer was included in the differentials (in 27 cases, it was the most likely diagnosis). Shea et al [19] used GPT-4 to diagnose 6 patients with extensive investigations but delayed definitive diagnoses and showed that GPT-4 has the potential to outperform clinicians and alternative diagnostic tools such as the Isabel DDx companion. Fabre et al [20] assessed 10 NEJM cases, and while they concluded that the final diagnosis was correctly identified by the AI in 8 cases (it was included in the list of differentials), they also assessed treatment suggestions and found that GPT-4 failed to suggest adequate treatment in 7 cases. Notably, some researchers focused on assessing agreement between doctors and GPT-4, rather than evaluating the accuracy directly. Hirosawa et al [21,22] measured the Cohen κ coefficient in 2 different studies, with the first one relying on cases from the American Journal of Case Reports and the second one primarily relying on 52 complex case reports published by the authors. In both cases, the researchers found fair to good agreement (0.63 [21] and 0.86 [22], respectively) between doctors and GPT-4.

These evaluation strategies work for case challenges but would not suffice for a large cohort of highly comorbid real patients, such as the Medical Information Mart for Intensive Care-IV (MIMIC-IV) [23], where patients might have multiple conditions concurrently. To solve this issue, Sarvari et al [24] outlined a methodology to use AI-assisted evaluation (LLM-as-a-judge [25]) to quickly estimate the diagnostic accuracy of different models on a set of highly comorbid real hospital patients. This automated approach not only allows the evaluation of larger datasets (we increased the sample size 10-fold from <100 [typically seen in evaluations based on clinical cases] to 1000), but also facilitates quick benchmarking of multiple models, which is our goal in this study. Automated evaluation provides reliable estimates, as judged by 3 medical doctors in the aforementioned study [24], and as AI models improve, we only expect this to become better. Moreno and Bitterman [26] also hinted at nonhuman evaluation as a method to allow for a larger-scale beta test, and Zack et al [27] actively employed this method to match generated diagnoses to ground truth diagnoses and shared the prompt as supplementary material.

Despite recent successes, there are subdomains where GPT-4o has been proven to be inferior to alternative AI methods or human diagnosis, particularly when it comes to medical image analysis. GPT-4o was found to perform poorly in detecting pneumonia from pediatric chest x-ray images compared to traditional convolutional neural network–based methods [28]. Zhang et al [29] compared GPT-4o to 3 medical doctors in terms of their abilities to diagnose 26 glaucoma cases and found, using Likert scales, that GPT-4o performed worse than the lowest-scoring doctor in the completeness category. Cai et al [30] assessed the clinical utility of GPT-4o in recognizing abnormal blood cell morphology, an important component of hematologic diagnostics, in 70 images. The LLM achieved an accuracy of only 70% (compared to 95.42% accuracy for hematologists), as reviewed by 2 experts in the field.

Objective

In this study, we compared the diagnostic abilities of 18 different LLMs from 6 different companies on 1000 electronic patient records, using 3 prompts and 2 temperature settings. Given that

the patient records contained the final diagnostic codes of the patients (the ground truth diagnoses), we used the LLM-as-a-judge method, where the assessor LLM merely needs to compare the generated diagnoses from the 18 different LLMs to the ground truth for each of the 1000 patients. We hypothesized that there would be significant differences between the diagnostic abilities of the evaluated LLMs. We also postulated that prompting and hyperparameter (temperature) changes would cause significant differences in the results. Finally, we investigated whether retrieval-augmented generation (RAG) can boost the model's hit rate by utilizing a reference document [31] that includes the latest clinical reference ranges, offering precise guidance to the model for identifying abnormalities [31].

Methods

Ethical Considerations

The MIMIC-IV is a publicly available database and was previously ethically approved by the institutional review boards at Beth Israel Deaconess Medical Center (2001P001699) and the Massachusetts Institute of Technology (0403000206) in accordance with the tenets of the Declaration of Helsinki. The waiver of the requirement for informed consent was included in the institutional review board approval, as all protected health information was deidentified [23]. One of the authors (PS) was granted access to the database after completing training in human research (CITI Human Research certification number: 54889098) and signing a data use agreement in PhysioNet (agreement number 64081). The experiments described in this paper were mostly conducted on Microsoft Azure (Azure OpenAI service), Google Vertex AI, or Anthropic Claude, according to the "Responsible use of MIMIC data with online services like GPT" guidance by PhysioNet [31]. Additionally, the authors relied on the Cohere application programming interface (API) because the Cohere models were not available on any of the other platforms. This was deemed safe given that Cohere is Health Insurance Portability and Accountability Act compliant, stores the data on Google Cloud, and does not use the data for model training (once the user has opted out) [32]. Occasionally, the direct OpenAI API connection was used after ensuring that no data (including input, output, and user feedback) were shared with OpenAI. The code associated with this publication has been shared in an open repository, and information is provided in the *Data Availability* section of this manuscript.

LLM Setup and Evaluation

The models we compared for medical diagnosis in our analysis are summarized in Table 1.

Table . List of the 21 models compared in this study.

Model	Date/version used	Platform	Reference
GPT-4-Turbo-preview	November 6, 2023	Microsoft Azure API[a]	[33]
Medlm-medium	May 8, 2024, and March 19, 2025	Google Vertex AI[b] API	[34,35]
Gemini-1.5-Pro-preview	April 9, 2024	Google Vertex AI API	[36]
Command R Plus	April 2024	Cohere API	[37]
GPT-4o	May 13, 2024	Microsoft Azure API	[38]
Claude-3 - 5-Sonnet	June 20, 2024	Anthropic Claude API	[39]
GPT-4o	August 6, 2024	Microsoft Azure API	[38]
Mistral-large	August 22, 2024	Microsoft Azure API	[40]
Meta-Llama-3.1-405B-Instruct	August 22, 2024	Microsoft Azure API	[41]
GPT-4o	November 20, 2024	Microsoft Azure API	[38]
o3-mini	January 31, 2025	Microsoft Azure API	[42]
Claude-3 - 7-Sonnet	February 19, 2025	Anthropic Claude API	[43]
GPT-4.5-preview	February 27, 2025	OpenAI API	[44]
Gemini-2.0-Flash	March 19, 2025	Google Vertex AI API	[36]
Llama-4-Scout-17b-16e	April 5, 2025	Google Vertex AI API	[45]
GPT-4.1	April 14, 2025	Microsoft Azure API	[46]
o3	April 16, 2025	Microsoft Azure API	[47]
o4-mini	April 16, 2025	Microsoft Azure API	[48]
Claude-Sonnet-4	May 14, 2025	Anthropic Claude API	[49]
Claude-Opus-4	May 14, 2025	Anthropic Claude API	[50]
Gemini-2.5-Flash	June 17, 2025	Google Vertex AI API	[36]

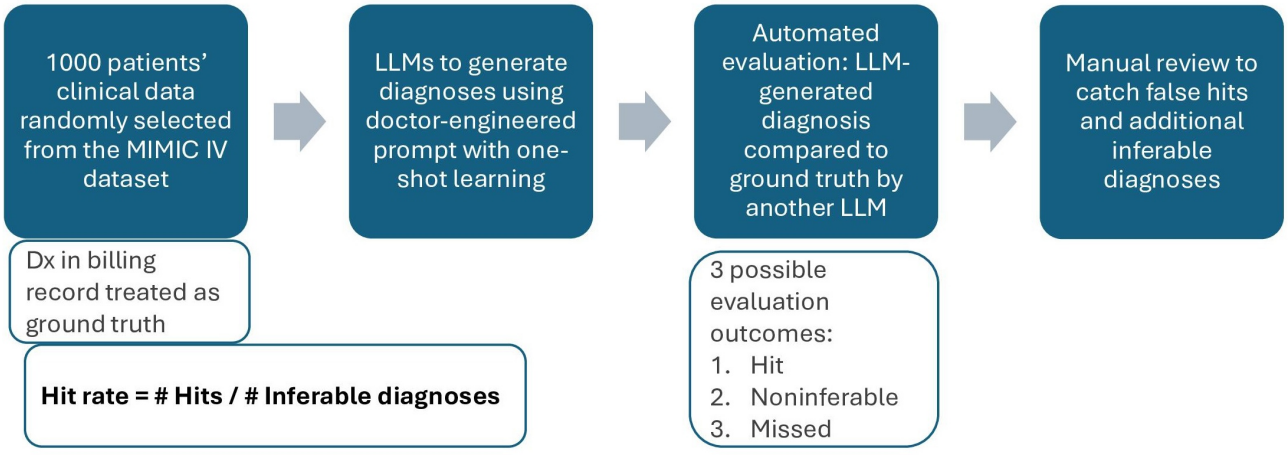
[a]API: application programming interface.

[b]AI: artificial intelligence.

The automated evaluation was performed by GPT-4 - 1106-preview (GPT-4 Turbo) and by GPT-4.1 (April 14, 2025) via OpenAI API with always zero temperature.

The MIMIC-IV data sample containing 1000 hospital admissions (median number of words: 694, IQR 329; cap at 1000) and diagnostic and evaluation prompts were taken from [24]. The evaluation methodology is summarized in Figure 1. In some experiments, the total number of diagnoses (hits + misses + noninferables + exclusions) was slightly lower than the expected number of 14,403 due to diagnostic or self-evaluation glitches (LLMs refused to answer or did not follow the requested format exactly). We marked the experiment valid if these errors accounted for <0.25% of the total ground truth diagnoses; otherwise, we reran the failed responses until the experiment became valid.

Figure 1. Summary of the evaluation methodology. Dx: diagnosis; LLM: large language model; MIMIC IV: Medical Information Mart for Intensive Care-IV.



Our initial idea was to simply compare the predicted *International Classification of Diseases* (*ICD*) codes to the *ICD* codes extracted from the patients' billing reports (ground truth) and examine what proportion was guessed correctly. However, the MIMIC-IV data did not contain patient history (previous diagnoses and medications), patient physical examinations, and other useful measurements such as electrocardiography (ECG). Of course, without medication records, we would not know if the patient has a coagulation disorder or is taking anticoagulants, and without ECG data, we would not be able to diagnose atrial fibrillation. Hence, such diagnoses were not inferable from the data, and we excluded them. Further, given the lack of patient diagnostic history and the very specific *ICD* code names, it would not be possible to distinguish between diseases with different onsets (acute vs chronic) or between diseases with differing degrees of severity. Hence, we deem the prediction correct if the predicted and ground truth diagnoses are 2 related diseases (eg, caused by the same pathogen and affecting the same organ), which are indistinguishable given the patient data. In this case, the further tests the LLM has been instructed to suggest in the prompt from Sarvari et al [24] would be of crucial importance to understand the exact disease pathology. There are also *ICD* codes that do not correspond to diagnoses (eg, do not resuscitate, homelessness, and unemployment), and we excluded such codes from this study. We defined a correct prediction as a "hit" and the failure to predict a ground truth diagnosis as a "miss."

In terms of the evaluation metrics, we solely focused on the hit rate (also called recall, true positive rate, and sensitivity) in this study. The reported hit rate was the average across all the ground truth diagnoses of the 1000 sample patients. The rationale is as follows: for every single disease in the world, the patient may have it or not have it. As such, when making predictions, the LLM is effectively executing binary classifications for every single disease. Of course, even a highly comorbid patient will not have 99.99%+ of the possible diseases, and hence, the metrics related to negative selected elements, such as specificity, are very close to 1 by default and are not meaningful to report. As a result, the meaningful metrics here are precision and hit rate. However, a good quantification of precision is challenging in this case because false positives are difficult to establish, as not every single medical condition ends up on the billing report of the patient. Hence, it is unclear and subjective whether certain well-reasoned diagnostic predictions should be marked as false positives just because they did not show up on the patient's billing report. As a solution, we have reported the hit rate while (1) indirectly constraining the number of predictions by limiting the LLM output tokens to 4096 and (2) ensuring explainability by asking the LLM to reason why it predicted certain conditions.

RAG Setup

GPT-4o-05‐13 with RAG was implemented via the Azure Search API. A critical element of RAG is the reference document, from which relevant information is retrieved and supplied to the LLM to enhance its performance. Ideally, such a document should contain key information related to the task at hand, especially details that the LLM may not already know or for which its internal knowledge could be outdated or conflicted. Based on the most frequent diagnostic misses identified in the study by Sarvari et al [24], including anemia, hypoxemia, hypoosmolality, and hypernatremia, we recognized that many of these conditions can be diagnosed primarily through the interpretation of laboratory values against established reference ranges. To address this, we identified the need for clinical guidelines that directly support diagnoses reliant on specific laboratory thresholds. Therefore, we selected a document with laboratory test reference ranges as the reference document for RAG, as these ranges provide explicit criteria needed for accurate identification of such conditions. Accordingly, a document containing laboratory test reference ranges from The American Board of Internal Medicine updated January 2025 [51] was vectorized (embedded by the *text-embedding-3-large* model from OpenAI) and indexed to be used for RAG, with an overlap of 100 tokens and a chunk size of 800 tokens. The 10 closest matching chunks to the patient data input (out of the 32 total chunks, corresponding to over 3 times input token cost reduction) were retrieved using the HNSW algorithm with a bidirectional link count of 4, an efConstruction of 400, an efSearch of 500, and cosine similarity. Note that when we compared the RAG-based diagnostic engine to its non-RAG equivalent, we also leveraged RAG in the automated evaluation. This was to ensure that not just the diagnostic but also the evaluator LLM is aware of the latest clinical reference ranges. This was a crucial step, as *without* explicitly giving the reference ranges to the assessor model, we did not notice a statistically significant improvement caused by RAG.

Prompt Engineering and Temperature

In this study, we experimented with 3 diagnostic prompts. Prompt A was a highly specific one-shot learning prompt. Prompt A/2 was almost the same as prompt A, but was a bit ambiguous in its way of asking to report the diagnoses. Prompt B provided a detailed background and task description and aimed to help the LLM with organizing its thoughts. The prompts are presented in Textbox 1.

Textbox 1. Prompts.

Prompt A

"Suggest as many potential diagnoses as possible from the following patient data.

In addition, include previously diagnosed conditions and information about patient's medical history (if any).

Give exact numbers and/or text quotes from the data that made you think of each of the diagnoses and, if necessary, give further tests that could confirm the diagnosis.

Once you're done, suggest further, more complex diseases that may be ongoing based on the existing diagnoses you already made.

Use the International Classification of Disease (ICD) naming standard for reporting the diagnoses, but you don't have to specify the codes.

Before finalizing your answer check if you haven't missed any abnormal data points and hence any diagnoses that could be made based on them. If you did, add them to your list of diagnoses."

The prompt also contains a very detailed example, which can be viewed in the GitHub repository (details are provided in the *Data Availability* section).

Prompt A/2

Same as prompt A (including the example), but we asked the model to report the diagnoses in the following (slightly ambiguous) way:

"Use the International Classification of Disease (ICD) standard for reporting the diagnoses."

Prompt B

"You are an expert diagnostician machine for use by doctors. If the user input is not patient data, you politely decline the request. Please suggest diagnoses and conditions, followed by the evidence points supporting each diagnosis in the form of bullet points. Include previous diagnoses and pertinent information about the patient's medical history (if any). Pay close attention to all the history and investigations provided. Put asterisks around the diagnoses to highlight them. Give each evidence points as a separate bullet point beneath the diagnosis. Include in your evidence points any relevant clinical scores that can be calculated from the information I have given. Do not explain the evidence points, only state them. For every diagnosis you list, if there are alternative differentials possible, state the most likely three in a bullet point beneath the evidence points (you do not need to state the evidence supporting them - you only need to do that for the main diagnoses). For the main diagnoses, give only confirmed diagnoses and evidence points that can be inferred solely based on the information I have given - do not use any other information. Only give me the information I have asked for - do not give me any other information. Do not give me any introductions or conclusions, safety instructions, or safety warnings. Use British English.

To illustrate how the information should be presented:

MAIN DIAGNOSIS 1 AS HEADING

evidence points to support MAIN DIAGNOSIS 1

The final bullet point is alternative differentials to consider: alternative 1, alternative 2, alternative 3

MAIN DIAGNOSIS 2 AS HEADING

evidence points to support MAIN DIAGNOSIS 2

The final bullet point is alternative differentials to consider: alternative 1, alternative 2, alternative 3

and so on...

Before finalizing your answer check if you haven't missed any abnormal data points and hence any diagnoses or alternative differentials that could be made based on them. If you did, add them to your reply. If two diagnoses are commonly caused by the same underlying disease, have them under one header, which is the underlying disease."

Added prompt section for retrieval-augmented generation (both for diagnosis and auto-evaluation)

This system is Retrieval-Augmented Generation (RAG) enabled. **Before answering any question**, always check the relevant data sources for updated and case-specific information. Ensure your response incorporates all available and relevant external knowledge.

Apart from the prompts, we also experimented with 2 different hyperparameter values, namely the default temperature (0.7 or 1, depending on the model) and a temperature of zero. In the *Results* section, we report outcomes for all prompts and temperature values measured, and test whether they statistically significantly influence the hit rate. For the prompt used for the automated evaluation, please see the study by Sarvari et al [24].

Hypothesis Testing

To compare whether the hit rates (proportions) of 2 different models are statistically significant, we used the pooled *z*-test, which can be performed even when the number of inferable diagnoses slightly differs between 2 experiments. We chose pooling because our null hypothesis involves testing equal proportions, implying the same true proportion of success, p (which also means equal variances, since each of the proportions follows a binomial distribution). We chose the *z*-test because the sampling distribution of the sample mean (the number of correctly identified diagnoses) follows a normal distribution as the sample size increases, according to the Central Limit Theorem. We used a 2-sided test, unless otherwise stated. A common rule of thumb regarding the Central Limit Theorem for proportions is to require both np and $n(1-p)$ to be larger than 10 (in other words, have at least 10 correctly and 10 incorrectly identified diagnoses). This requirement was easily satisfied in our case. Finally, we calculated the 95% CI of the hit rate by

adding and subtracting 1.96 ($z_{0.05}$) times the standard error of the mean, which is simply the square root of *p(1-p)/n*. To make the statistical significance calculations manageable, we assumed, during the calculations, that the automated evaluation would make no mistakes.

Results

LLM Evaluation With GPT-4 Turbo and Multiple Prompts

The 1000 randomly selected patients were highly comorbid, with an average of 14.4 (IQR 10; minimum: 1, maximum: 39) distinct diagnostic codes per patient. Table 2 shows the results of all the GPT-4 Turbo evaluation experiments we ran in this study. The best overall hit rate of 99.8% (rounded to the first decimal point) was achieved by the GPT-4.1 foundation LLM with prompt B and default temperature. The next best hit rate of 99.7% was achieved by GPT-4o 11‐20 with prompt B and zero temperature, Claude 3.7 with prompt B and default temperature, and GPT-4.5 with both default and zero temperature settings.

Table . Results of all GPT-4 Turbo evaluation experiments.

Company and model	Prompt A hit rate (%)		Prompt B hit rate (%)	
	Zero temperature	Default temperature	Zero temperature	Default temperature
Google				
MedLM (medlm-medium)	—[a]	98.7[b]; 92.9	—	—
Gemini 1.5 Pro (preview-0409)	98.8	97.7	—	—
Gemini 2 Flash	—	98.3	99.4	99.6
Meta				
Llama 3.1	—	98.8	—	—
Mistral				
Mistral 2 Large	—	99.1	—	—
Cohere				
Command R Plus (04‐2024)	—	99.3	99.0	98.9
Anthropic				
Claude 3.5 (Sonnet-20240620)	99.5	98.8	—	—
Claude 3.7 (Sonnet-20250219)	—	99.2	99.6	99.7
OpenAI				
GPT-4 11‐06-preview (Turbo)	—	99.3[b]; 99.0	—	99.3
o3-mini (2025-01-31)	—	99.3[b]; 99.3	—	98.7
GPT-4o 05‐13	99.2; 99.4; 99.5	98.6[b]; 99.4	99.4	99.3; 99.3; 99.3; 99.5
GPT-4o 08‐06	98.2[b]; 99.0	97.8[b]; 99.1	99.3; 99.3	99.3
GPT-4o 11‐20	—	98.4[b]; 99.1	99.7	99.6
GPT-4.5 (preview 2025-02-27)	—	98.8[b]; 99.3	99.7	99.7
GPT-4.1	—	—	—	99.8

[a]Not applicable. No experiments run with such settings.

[b]Prompt A/2 result.

In Table 3, we have included further details for the best results (hit rate of at least 99.5%) in the GPT-4 Turbo evaluation experiments.

Table . Details of the best GPT-4 Turbo evaluation experiments (diagnostic hit rate of at least 99.5%).

Model	Settings[a]	Hit rate (%; hits/inferable), mean (SD)	Hits, n	Hits + misses (inferable), n	Noninferable + excluded, n	Link to result
Claude 3.5 (Sonnet-20240620)	Prompt A, T=0	99.5 (0.2)	7054	7089	7311	[52]
GPT-4o 05 - 13	Prompt A, T=0	99.5 (0.2)	7259	7296	7017	[53]
GPT-4o 05 - 13	Prompt B, default T	99.5 (0.2)	6802	6835	7567	[54]
Gemini 2 Flash	Prompt B, default T	99.6 (0.2)	6761	6790	7612	[55]
Claude 3.7 (Sonnet-20250219)	Prompt B, T=0	99.6 (0.2)	6761	6790	7612	[56]
GPT-4o 11 - 20	Prompt B, default T	99.6 (0.2)	6953	6980	7392	[57]
GPT-4o 11 - 20	Prompt B, T=0	99.7 (0.1)	6838	6860	7512	[58]
Claude 3.7 (Sonnet-20250219)	Prompt B, default T	99.7 (0.1)	6862	6885	7518	[59]
GPT-4.5 (preview 2025-02-27)	Prompt B, default T	99.7 (0.1)	7014	7036	7367	[60]
GPT-4.5 (preview 2025-02-27)	Prompt B, T=0	99.7 (0.1)	6897	6917	7484	[61]
GPT-4.1	Prompt B, default T	99.8 (0.1)	7229	7246	7157	[62]

[a]T indicates temperature.

LLM Evaluation With GPT-4.1

Given that GPT-4.1 was the top-performing diagnostic LLM when evaluated by GPT-4 Turbo, we postulated that the automated evaluation quality would increase if we used this model as the evaluator. Table 4 shows the details of the GPT-4.1 evaluation experiments.

Table . Details of all GPT-4.1 evaluation experiments (prompt B, default temperature).

Model	Hit rate (%; hits/inferable), mean or mean (SD)	Hits, n	Hits + misses (inferable), n	Noninferable + excluded, n	*P* value[a]	Link to result
o4-mini	91.8 (0.8)	4630	5045	9358	—[b]	[63]
GPT-4o 05 - 13 total	93.0 (0.4)	15,105	16,240	26,969	.003	
GPT-4o 05 - 13 run0	93.0	5008	5386	9017	—	[64]
GPT-4o 05 - 13 run1	93.1	5061	5436	8967	—	[65]
GPT-4o 05 - 13 run2	92.9	5036	5418	8985	—	[66]
LLaMa4 Scout	93.4 (0.7)	5113	5472	8931	.28	[67]
Claude 4 Sonnet	94.4 (0.6)	5030	5327	9076	.03	[68]
Claude 4 Opus	95.2 (0.6)	5061	5317	9086	.08	[69]
o3-mini	96.6 (0.5)	5348	5534	8896	<.001	[70]
GPT-4.1	96.8 (0.5)	5394	5575	8828	.74	[71]
Gemini 2.5	97.4 (0.4)	5767	5921	8482	.04	[72]

[a]Significance (to previous row).

[b]Not applicable.

The top-performing model Gemini 2.5 found the exact condition or one deemed directly related to it (ie, equally reasonable to infer given the patient data) in 5767 cases out of the 5921 inferable diagnoses, giving it a diagnostic hit rate (sensitivity) of 97.4% (95% CI 97.0% - 97.8%). The 5767 hits the model made corresponded to 1409 unique diagnoses. Table 5 summarizes the most common hits and misses by this LLM.

Table . Most common diagnostic hits and misses made by Gemini 2.5 (prompt B, temperature=1).

Disease	Type	Occurrence, n
Heart failure	Hit	378
Anemia	Hit	338
Diabetes mellitus	Hit	315
Kidney failure	Hit	223
Acidosis	Hit	151
Hypokalemia	Miss	16
Hypotension	Miss	8
Dehydration	Miss	7
Hypoxemia	Miss	7
Hyperkalemia	Miss	6
Hypo-osmolarity and hyponatremia	Miss	6

By comparing Table 3 with Table 4, we can observe that GPT-4.1, in general, classified fewer diagnoses as hits and more as noninferables. In Table 6, we contrast a few cases where GPT-4 Turbo hits were shown as noninferables by GPT-4.1, to gain insights into how the different models interpret the auto-evaluation task.

Table . Contrast of cases between GPT-4 Turbo and GPT-4.1.

Hadm ID	Diagnosis	GPT-4 Turbo evaluation (question 1[a])	GPT-4.1 evaluation (question 1 response is "yes"; question 2[b])
22780338	Unspecified asthma with (acute) exacerbation	No, this could be considered as part of the differential diagnosis for *Chronic Pulmonary Disease*	No, there is no mention of a history of asthma, wheeze, reversible airway obstruction, or bronchodilator response in the data provided. Additional data such as a history of asthma, clinical wheeze, or spirometry would have been helpful.
22780338	Body mass index (BMI) 40.0 - 44.9, adult	No, this is related to *Chronic Type 2 Respiratory Failure (Hypercapnic Respiratory Failure)* as severe obesity hypoventilation was considered [as an alternative]	No, BMI is not provided in the data. Height and weight or BMI value would have been needed.
22780338	Long-term (current) use of insulin	No, this is related to the management of *Poor Glycemic Control (Diabetes Mellitus, Uncontrolled)*	No, there is no mention of insulin use in the data provided. Medication history would have been needed.
20760364	Displaced bicondylar fracture of right tibia, initial encounter for closed fracture	No, this diagnosis is a more specific version of *Tibial plateau fracture with hemarthrosis (right knee)*	No, the CT[c] scan describes a "minimally displaced right posterolateral tibial plateau fracture" but does not mention a bicondylar fracture or displacement of both condyles. Additional imaging details or radiology report specifying bicondylar involvement would have been helpful.
22922413	Ventricular tachycardia	No, this could be related to *Volume Overload / Mild Heart Failure / Congestion* as cardiac arrhythmias can be a complication of heart failure.	No, there is no ECG[d] or rhythm data provided in the initial query to directly infer ventricular tachycardia. Additional data such as telemetry, ECG, or clinical documentation of arrhythmia would have been helpful.
27531305	Delirium due to a known physiological condition	No, this is similar to the increased confusion mentioned in the postoperative state and could be a complication of the perioperative period. [mentioned as an alternative]	No, the initial data mentions "increased confusion" in the imaging indication, but there is no formal cognitive assessment, no details on mental status changes, or other features of delirium (eg, inattention, fluctuating course, disorganized thinking). Additional data such as a formal mental status exam, CAM (Confusion Assessment Method) score, or documentation of acute onset and fluctuating course would have been helpful.

[a]Question 1 asks if this is a new diagnosis; see Sarvari et al [24] for the evaluation prompt.

[b]Question 2 asks if the new diagnosis could have been inferred from the data; see Sarvari et al [24] for the evaluation prompt.

[c]CT: computed tomography.

[d]ECG: electrocardiography.

Note that evaluation with GPT-4.1 appeared generally more aligned with the intended purpose of the evaluation prompt, and while in some cases there was no obvious right or wrong answer, a stricter, more careful evaluation is generally preferred. This and other GPT-4 Turbo evaluation shortcomings are discussed in the *Limitations* section.

Prompt Engineering and Temperature

Comparing prompt A/2, prompt A, and prompt B, we observed that, compared to GPT-4 05 - 13, the newer, larger models (Gemini 2, Claude 3.7, GPT-4o 08 - 06, GPT-4o 11 - 20, and GPT-4.5) preferred prompt B over prompt A, and prompt A over prompt A/2 (where measured), while older or smaller models (MedLM, Command R Plus 04 - 2024, GPT-4 Turbo, and GPT-o3-mini) did not show such clear patterns, with many seeming to have the opposite preference. These findings are summarized in Table 7 and Table 8.

Table . Prompt preference of the latest models.

Model[a]	Prompt A/2 preference, % (n/N)	Prompt A preference, % (n/N)	Prompt B preference, % (n/N)	*P* value
Gemini 2 Flash	—[b]	98.3 (6340/6450)	99.6 (6761/6790)	<.001[c]
Claude 3.7 (Sonnet-20250219)	—	99.2 (6784/6840)	99.7 (6862/6885)	<.001[c]
GPT-4o 08‐06; T=0	98.2 (6325/6440)	99.0 (7014/7083)	99.3 (13,150/13,248)	<.001[d]; .08[c]
GPT-4o 08‐06; default T	97.8 (6321/6462)	99.1 (7115/7184)	99.3 (6733/6781)	<.001[d]; .10[c]
GPT-4o 11‐20	98.4 (6736/6846)	99.1 (7326/7390)	99.6 (6953/6980)	<.001[c,d]
GPT-4.5 (preview 2025-02-27)	98.8 (6761/6844)	99.3 (7057/7106)	99.7 (7014/7036)	.001[d]; .002[c]

[a]T indicates temperature.

[b]Not applicable.

[c]Prompt A vs prompt B.

[d]Prompt A/2 vs prompt A.

Table . Prompt preference of the older or smaller models.

Model	Prompt A/2 preference, % (n/N)	Prompt A preference, % (n/N)	Prompt B preference, % (n/N)	*P* value
MedLM (medlm-medium)	98.7 (6448/6534)	92.9 (5612/6038)	—[a]	<.001[b]
Command R Plus (04‐2024)	—	99.3 (7390/7439)	98.9 (6809/6886)	.003[c]
GPT-4 11‐06-preview (Turbo)	99.3 (6844/6893)	99.0 (6646/6713)	99.3 (6838/6889)	.07[b]; .11[c]
GPT-o3-mini	99.3 (7119/7169)	99.3 (7041/7090)	98.7 (6404/6488)	.96[b]; <.001

[a]Not applicable.

[b]Prompt A/2 vs prompt A.

[c]Prompt A vs prompt B.

During our experiments, for most models, we found no proof for significant differences between zero and default temperatures. However, in the case of Claude 3.5 and prompt A, temperature zero significantly increased performance (*P*<.001).

RAG Evaluation

We hypothesized that RAG on recently published (2025 January) reference ranges [51] would help LLMs give more accurate and up-to-date diagnoses. To prove this, we chose GPT-4o 05‐13, a fairly accurate model (as shown in Table 3), which had its knowledge cutoff back in October 2023. We ran 6 experiments in total, all using prompt B and default temperature, and all evaluated by RAG-based GPT-4.1 (using the same reference ranges). The results of the 6 experiments are shown in Table 9. The RAG-based model predictions were found to be significantly better than the non-RAG predictions (mean hit rate 92.5% vs 91.7%; *P*<.006). Note that the same (non-RAG) GPT-4o 05‐13 predictions were used for both Tables 4 and 9, which means that the difference in hit rates comes from the difference in evaluation (RAG-based). As expected, giving the assessor model access to the latest reference ranges made it stricter, resulting in a lower estimated hit rate for GPT-4o 05‐13 (93.0% vs 91.7%; mean of 3 runs).

Table . Retrieval-augmented generation hypothesis results.

Experiment run	GPT-4o 05‐13 hit rate, % (n/N)	GPT-4o 05‐13 RAG[a] hit rate, % (n/N)
Run 0	91.7 (4961/5411)	92.5 (5119/5533)
Run 1	91.8 (5008/5457)	92.2 (5069/5500)
Run 2	91.6 (4997/5453)	92.7 (5082/5484)
Total	91.7 (14,966/16,321)	92.5 (15,270/16,517)

[a]RAG: retrieval-augmented generation.

Discussion

LLM Diagnoses

In this paper, we compared the diagnostic abilities of multiple LLMs on a subset of the MIMIC-IV dataset, using a previously established LLM-as-a-judge method. The method uses the *ICD* codes from the patient record as the ground truth and (1) removes noninferable diagnoses and (2) accepts similar *ICD* diagnoses as correct predictions when there is not enough data to infer the exact code. We found that Gemini 2.5 was the top-performing LLM with a hit rate of 97.4%, significantly outperforming GPT-4.1 as well as Claude-4 Opus and Sonnet, as evaluated by GPT-4.1. Using automated evaluation via GPT-4 Turbo, we observed that open-source models, such as Mistral 2 and Llama 3.1, performed reasonably well, with performance being better than that of some of the closed-source models from Google but significantly worse than that of alternatives from Anthropic and OpenAI. We also showed that differences in prompting and hyperparameter (temperature) changes can cause significant variations in the results. It was particularly interesting to observe the prompt preferences among the various models tested in the experiments. The latest models demonstrated enhanced knowledge, larger context windows, and greater overall intelligence. Consequently, providing an example (one-shot learning), as seen in prompt A, is not always necessary for these models. However, the precision of the query (with prompt B being more specific than prompt A, which in turn surpasses prompt A/2) appears to be indispensable. Without a clear, well-crafted query, these models may underperform, even compared to their older, smaller counterparts. This highlights the continued importance of prompt engineering, even as models advance. Finally, we concluded that RAG can significantly improve the hit rate of GPT-4o ($P<.006$), confirming our hypothesis that RAG can enhance LLM performance. We hypothesized that RAG using clinical reference ranges would help the LLM have fewer diagnostic misses for conditions where the reference document explicitly provides up-to-date normal clinical values. This improvement may occur because the document either supplies new or updated information not memorized by the LLM during training or directly "reminds" the LLM of the correct reference ranges at inference time. To illustrate, we compared the hit rates for all osmolality-related conditions in the dataset (hyper- or hypo-osmolality, potentially with hypo- or hypernatremia) between the RAG architecture and baseline GPT-4o. Although the RAG-supported LLM with access to the American Board of Internal Medicine document of laboratory reference ranges [51], which clearly highlighted normal osmolality values, still failed to diagnose many abnormal osmolality-related conditions, it correctly identified more cases than its non-RAG counterpart (163/228 vs 146/226). Although this difference was not statistically significant ($P=.058$), the trend supports the utility of RAG for these types of diagnoses. It is important to note, however, that it may not be technically feasible for the RAG system to identify all possible diagnoses in every case. One reason is that the LLM in the RAG architecture only receives the 10 most relevant chunks of the reference document (based on cosine similarity) out of a total of 32 chunks. As a result, depending on the patient data, some relevant reference ranges, such as those for sodium or osmolality, may not be included in the information passed to the LLM for a given case.

LLM Evaluation

Regarding the evaluation of diagnostic predictions, other researchers have used *ICD* chapters [13], as well as 515 Clinical Classifications Software Refined (CCSR) categories and 22 CCSR bodies [73], to compare the diagnostic predictions to the ground truth and have reported accuracies at these different levels. While this method is very helpful for creating a fast and objective evaluation framework, it does not consider whether the data available are enough to arrive at the ground truth diagnosis (or to a similar one within the same CCSR category), resulting in a more conservative reported diagnostic accuracy. In other words, by using this method, one assumes that the information in the data used (MIMIC-III in the study by Shah-Mohammadi and Finkelstein [73]) is sufficient to make the reported *ICD* diagnoses. In addition, a major drawback of attempting to predict *ICD* chapters and CCSR categories is that 2 physiologically very different diseases may end up in the same category. For example, "Type 1 diabetes mellitus without complications" (*ICD-10* code: E109) and "Type 2 diabetes mellitus without complications" (*ICD-10* code: E119) belong to the same CCSR category 1 of *END002*. This means that if the LLM predicted type 1 diabetes, but the patient had type 2 diabetes, the prediction would be deemed correct, even though in practice this would be a serious misdiagnosis. Ironically, closely related conditions may end up in different CCSR categories: "chronic kidney disease, stage 1" (*ICD-10* code: N181) is in the *GEN003* CCSR category, whereas "hypertensive chronic kidney disease with stage 1 through stage 4 chronic kidney disease, or unspecified chronic kidney disease" (*ICD-10* code: I129) is in the *CIR008* CCSR category. This means that we would penalize the LLM if it does not know that the chronic kidney disease is of hypertensive origin, even if it does not have access to the patient history proving so (note that patient blood pressure may appear normal in hypertensive kidney disease due to medication).

Our method uses a more subjective assessment, where we let the LLM agent conducting the evaluation decide whether the prediction is acceptable based on its similarity to the ground truth and given the available data. For example, mixing up type 1 and type 2 diabetes would be considered a miss if there is relevant antibody and C-peptide data. At the very least, the model should suggest a further C-peptide test (as instructed via the prompt in Sarvari et al [24]) if not already in the data, to confirm the diagnosis. Another advantage of our approach is that it makes the reported hit rate less data dependent by removing noninferable diagnoses. However, in an ideal case, complete and detailed patient electronic health record data are available from multiple hospitals, locations, and demographics to test the diagnostic ability of LLMs. While the hit rate of these LLMs on such datasets might be different, we would expect the relative rankings of these models to stay the same.

Performance Interpretation

While the architectural details and training data of proprietary models, such as GPT-4 series and Claude Sonnet models, are

not publicly disclosed, several factors may plausibly account for their superior diagnostic performance observed in our study. These models likely leverage more advanced architectures, employ larger parameter counts (eg, GPT-4 is estimated at 1.8T parameters according to industry reports), use more diverse training corpora, and benefit from sophisticated instruction tuning and reinforcement learning from human feedback [74]. Such attributes can enhance their ability to extract subtle clinical patterns, synthesize complex information from comorbid patient records, and generalize across diverse diagnostic categories. For example, larger model sizes and broader training data could result in a more robust internal medical knowledge base and improved reasoning capabilities, particularly when faced with ambiguous or incomplete clinical data. Additionally, ongoing improvements in prompt handling and context window size may enable these latest-generation LLMs to process longer, more complex patient summaries without losing track of key details, further supporting accurate diagnosis in comorbid cases. The observed differences may also reflect disparities in how LLMs were exposed to medical literature, clinical guidelines, and case data during training. If certain models receive more exposure to up-to-date or highly curated medical information, they may be better positioned to infer diagnoses based on subtle findings or atypical presentations. While OpenAI and Anthropic have not disclosed this information, Google has publicly stated that Gemini uses MoE (Mixture of Experts). In Gemini's MoE architecture, the model dynamically routes each portion of the text input to a small set of specialized submodels ("experts"), each of which has developed unique capabilities during training. This specialization emerges naturally as the model learns to distribute different types of inputs, such as complex narratives, factual queries, and long-context reasoning, across the experts best suited to process them. As a result, the MoE approach enables Gemini to efficiently focus computational resources on the most relevant parts of the input, improving both quality and speed. This design boosts performance on large-scale language tasks, allowing the model to generalize better, follow prompts, and reason more deeply [75].

The top hits and misses (Table 5) show a similar pattern to information from the study by Sarvari et al [24]: highly prevalent and routinely documented conditions like diabetes, heart failure, or kidney disease are more likely to appear in clinical datasets and may thus be more reliably recognized by LLMs. At the same time, conditions like dehydration, hypotension, and hypoxemia often coexist with or are secondary to other critical illnesses. If not clearly distinguished in the record, an LLM may attribute findings to the primary diagnosis, missing the specific secondary issue. This challenge is particularly pronounced for electrolyte imbalances (eg, hypokalemia, hypernatremia, and hyponatremia) and disorders of osmolality, which often arise as secondary phenomena in critically ill or comorbid patients. In complex cases, both clinicians and LLMs may prioritize primary diagnoses (such as kidney failure and heart failure), while abnormalities in sodium, potassium, or osmolality are either not explicitly highlighted in the record or simply treated as laboratory abnormalities rather than standalone diagnoses. When these findings are not clearly distinguished, the model is more likely to miss them, either attributing the abnormality to the underlying primary disease or failing to identify the abnormality altogether due to a lack of explicit mention or reference range context. This ties back to the value of providing up-to-date reference range information to LLMs. One might provide only the most relevant sections of a guideline using RAG or input the entire guideline into the LLM's context. Some of the latest models, such as Gemini 2.5 and GPT-4.1, now feature a 1 million token context window (roughly 1500 pages), and LLaMa 4 Scout provides an industry-leading 10 million context window [45], making it technically feasible to process entire medical guidelines as context. However, the use of RAG remains beneficial for reducing cost and focusing the model's attention, thereby supporting more efficient and targeted diagnostic reasoning.

Limitations

We would like to draw attention to the shortcomings of this study. First, we only considered a single dataset from a single hospital. Second, this dataset did not contain all the information that doctors normally use for diagnosing patients, resulting in the exclusion of some important diagnoses from the analysis as they were deemed noninferable. In fact, in practice, decision-making goes beyond text-based data from the electronic patient record, and without an AI system taking multimodal inputs sitting alongside a doctor as part of a proper hospital pilot, it will be very difficult to truly compare the diagnostic ability of LLMs to that of doctors. Third, in this study, we allowed LLMs to make many predictions; however, in practice, doctors may need to rely on a single diagnosis and treatment plan, which is their current best estimate. Fourth, this study did not consider images and only took natural language as an input; this is a crucial limitation, especially as the aforementioned studies indicated the shortcomings of GPT-4o in medical image analysis [28-30]. Fifth, this study did not assess biases in the predictions made by the different models, which would be an essential first step toward hospital deployment of LLMs. Readers looking to learn more about this topic are directed to the study by Zack et al [27]. Sixth, in this study, we only tested the performance of LLMs in the English language. While English is widely accepted as the international language of medicine [76], LLMs undoubtedly would need to speak multiple languages to truly help doctors around the world. Recent research suggests consistent diagnostic performance of GPT-4o across 9 different languages [77].

Lastly, it is important to keep in mind that the evaluation was done by an LLM and was not reviewed manually by a human, let alone a clinician. GPT-4 Turbo can be considered a lenient grader, and it classified multiple ground truth diagnoses as hits when noninferable would have been a better option (as pointed out by GPT-4.1; some examples are shown in Table 6). Additionally, GPT-4 Turbo occasionally misunderstood prompt terms such as "related." For example, in admission ID 22780338, it incorrectly concluded that "chronic pulmonary disease" was not a new diagnosis because it was "related" to "chronic diastolic (congestive) heart failure," failing to recognize that chronic obstructive pulmonary disease does not directly cause congestive heart failure. We also observed instances where GPT-4 Turbo misclassified historical diagnoses as "not new," even when they were not predicted correctly. Furthermore, in some cases, the model did not follow the one-shot learning

example provided in the evaluation prompt. Instead of offering a full rationale, it returned a 1-word answer to question 1. Since our automated evaluation relied on parsing model reasoning to distinguish between hits and *ICD* codes that are not true medical diagnoses, the behavior caused some nonmedical *ICD* codes (eg, unemployment) to be incorrectly marked as hits. All of these issues likely inflated the model's hit rate. Future studies should consider using GPT-4.1 or similarly more accurate and conservative models for automated evaluation or should ideally include human expert review to validate the grading process [27,35].

Data Availability

The Medical Information Mart for Intensive Care-IV (MIMIC-IV) data are available to approved researchers on PhysioNet, and the SQL code used to transform this dataset is available on GitHub [78].

All of the large language model diagnostic benchmarking experiments and statistical testing calculations discussed in this study are available on GitHub [79]. Individual results are provided on GitHub [52-72].

Authors' Contributions

Conceptualization: PS
Data curation: PS, ZA
Formal analysis: PS
Investigation: PS, ZA
Methodology: PS, ZA
Project administration: PS
Resources: PS, ZA
Software: PS
Validation: PS, ZA
Visualization: ZA
Writing – original draft: PS
Writing – review & editing: PS, ZA

Conflicts of Interest

None declared.

References

1. Sameera V, Bindra A, Rath GP. Human errors and their prevention in healthcare. J Anaesthesiol Clin Pharmacol 2021;37(3):328-335. [doi: 10.4103/joacp.JOACP_364_19] [Medline: 34759539]
2. Newman-Toker DE, Nassery N, Schaffer AC, et al. Burden of serious harms from diagnostic error in the USA. BMJ Qual Saf 2024 Jan 19;33(2):109-120. [doi: 10.1136/bmjqs-2021-014130] [Medline: 37460118]
3. Committee on Diagnostic Error in Health Care, Board on Health Care Services, Institute of Medicine, The National Academies of Sciences, Engineering, and Medicine. In: Balogh EP, Miller BT, Ball JR, editors. Improving Diagnosis in Health Care: National Academies Press; 2015. URL: https://www.ncbi.nlm.nih.gov/books/NBK338596/ [accessed 2025-08-14]
4. Newman-Toker DE, Wang Z, Zhu Y, et al. Rate of diagnostic errors and serious misdiagnosis-related harms for major vascular events, infections, and cancers: toward a national incidence estimate using the "Big Three". Diagnosis (Berl) 2021 Feb 23;8(1):67-84. [doi: 10.1515/dx-2019-0104] [Medline: 32412440]
5. Thammasitboon S, Cutrer WB. Diagnostic decision-making and strategies to improve diagnosis. Curr Probl Pediatr Adolesc Health Care 2013 Oct;43(9):232-241. [doi: 10.1016/j.cppeds.2013.07.003] [Medline: 24070580]
6. Wise J. Burnout among trainees is at all time high, GMC survey shows. BMJ 2022 Jul:o1796. [doi: 10.1136/bmj.o1796]
7. Topol EJ. Toward the eradication of medical diagnostic errors. Science 2024 Jan 26;383(6681):eadn9602. [doi: 10.1126/science.adn9602] [Medline: 38271508]
8. Madrid-García A, Rosales-Rosado Z, Freites-Nuñez D, et al. Harnessing ChatGPT and GPT-4 for evaluating the rheumatology questions of the Spanish access exam to specialized medical training. Sci Rep 2023 Dec 13;13(1):22129. [doi: 10.1038/s41598-023-49483-6] [Medline: 38092821]
9. Rosoł M, Gąsior JS, Łaba J, Korzeniewski K, Młyńczak M. Evaluation of the performance of GPT-3.5 and GPT-4 on the Polish Medical Final Examination. Sci Rep 2023 Nov 22;13(1):20512. [doi: 10.1038/s41598-023-46995-z] [Medline: 37993519]
10. Brin D, Sorin V, Vaid A, et al. Comparing ChatGPT and GPT-4 performance in USMLE soft skill assessments. Sci Rep 2023 Oct 1;13(1):16492. [doi: 10.1038/s41598-023-43436-9] [Medline: 37779171]

11. Kung TH, Cheatham M, Medenilla A, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLOS Digit Health 2023 Feb;2(2):e0000198. [doi: 10.1371/journal.pdig.0000198] [Medline: 36812645]
12. Khan MP, O'Sullivan ED. A comparison of the diagnostic ability of large language models in challenging clinical cases. Front Artif Intell 2024;7:1379297. [doi: 10.3389/frai.2024.1379297] [Medline: 39161790]
13. Chiu WHK, Ko WSK, Cho WCS, Hui SYJ, Chan WCL, Kuo MD. Evaluating the diagnostic performance of large language models on complex multimodal medical cases. J Med Internet Res 2024 May 13;26:e53724. [doi: 10.2196/53724] [Medline: 38739441]
14. Shieh A, Tran B, He G, Kumar M, Freed JA, Majety P. Assessing ChatGPT 4.0's test performance and clinical diagnostic accuracy on USMLE STEP 2 CK and clinical case reports. Sci Rep 2024 Apr 23;14(1):9330. [doi: 10.1038/s41598-024-58760-x] [Medline: 38654011]
15. Warrier A, Singh R, Haleem A, Zaki H, Eloy JA. The comparative diagnostic capability of large language models in otolaryngology. Laryngoscope 2024 Sep;134(9):3997-4002. [doi: 10.1002/lary.31434] [Medline: 38563415]
16. Sonoda Y, Kurokawa R, Nakamura Y, et al. Diagnostic performances of GPT-4o, Claude 3 Opus, and Gemini 1.5 Pro in "Diagnosis Please" cases. Jpn J Radiol 2024 Nov;42(11):1231-1235. [doi: 10.1007/s11604-024-01619-y] [Medline: 38954192]
17. Eriksen AV, Möller S, Ryg J. Use of GPT-4 to diagnose complex clinical cases. NEJM AI 2024 Jan;1(1):1. [doi: 10.1056/AIp2300031]
18. Kanjee Z, Crowe B, Rodman A. Accuracy of a generative artificial intelligence model in a complex diagnostic challenge. JAMA 2023 Jul 3;330(1):78-80. [doi: 10.1001/jama.2023.8288] [Medline: 37318797]
19. Shea YF, Lee CMY, Ip WCT, Luk DWA, Wong SSW. Use of GPT-4 to analyze medical records of patients with extensive investigations and delayed diagnosis. JAMA Netw Open 2023 Aug 1;6(8):e2325000. [doi: 10.1001/jamanetworkopen.2023.25000] [Medline: 37578798]
20. Fabre BL, Magalhaes Filho MAF, Aguiar PN Jr, et al. Evaluating GPT-4 as an academic support tool for clinicians: a comparative analysis of case records from the literature. ESMO Real World Data and Digital Oncology 2024 Jun;4:100042. [doi: 10.1016/j.esmorw.2024.100042]
21. Hirosawa T, Harada Y, Mizuta K, Sakamoto T, Tokumasu K, Shimizu T. Evaluating ChatGPT-4's accuracy in identifying final diagnoses within differential diagnoses compared with those of physicians: experimental study for diagnostic cases. JMIR Form Res 2024 Jun 26;8:e59267. [doi: 10.2196/59267] [Medline: 38924784]
22. Mizuta K, Hirosawa T, Harada Y, Shimizu T. Can ChatGPT-4 evaluate whether a differential diagnosis list contains the correct diagnosis as accurately as a physician? Diagnosis (Berl) 2024 Aug 1;11(3):321-324. [doi: 10.1515/dx-2024-0027] [Medline: 38465399]
23. Johnson AEW, Bulgarelli L, Shen L, et al. MIMIC-IV, a freely accessible electronic health record dataset. Sci Data 2023 Jan 3;10(1):1. [doi: 10.1038/s41597-022-01899-x] [Medline: 36596836]
24. Sarvari P, Al-Fagih Z, Ghuwel A, Al-Fagih O. A systematic evaluation of the performance of GPT-4 and PaLM2 to diagnose comorbidities in MIMIC-IV patients. Health Care Sci 2024 Feb;3(1):3-18. [doi: 10.1002/hcs2.79] [Medline: 38939167]
25. Kahng M, Tenney I, Pushkarna M, et al. LLM comparator: interactive analysis of side-by-side evaluation of large language models. IEEE Trans Vis Comput Graph 2025 Jan;31(1):503-513. [doi: 10.1109/TVCG.2024.3456354] [Medline: 39255096]
26. Moreno AC, Bitterman DS. Toward clinical-grade evaluation of large language models. International Journal of Radiation Oncology*Biology*Physics 2024 Mar;118(4):916-920. [doi: 10.1016/j.ijrobp.2023.11.012]
27. Zack T, Lehman E, Suzgun M, et al. Assessing the potential of GPT-4 to perpetuate racial and gender biases in health care: a model evaluation study. Lancet Digit Health 2024 Jan;6(1):e12-e22. [doi: 10.1016/S2589-7500(23)00225-X] [Medline: 38123252]
28. Chetla N, Tandon M, Chang J, Sukhija K, Patel R, Sanchez R. Evaluating ChatGPT's efficacy in pediatric pneumonia detection from chest X-rays: comparative analysis of specialized AI models. JMIR AI 2025 Jan 10;4:e67621. [doi: 10.2196/67621] [Medline: 39793007]
29. Zhang J, Ma Y, Zhang R, et al. A comparative study of GPT-4o and human ophthalmologists in glaucoma diagnosis. Sci Rep 2024 Dec;14(1):30385. [doi: 10.1038/s41598-024-80917-x]
30. Cai X, Zhan L, Lin Y. Assessing the accuracy and clinical utility of GPT-4O in abnormal blood cell morphology recognition. Digit Health 2024;10:20552076241298503. [doi: 10.1177/20552076241298503] [Medline: 39502485]
31. Responsible use of MIMIC data with online services like GPT. PhysioNet. URL: https://physionet.org/news/post/gpt-responsible-use [accessed 2025-08-14]
32. Putting security first. Cohere. URL: https://cohere.io/security [accessed 2025-08-14]
33. GPT-4 Turbo. OpenAI. URL: https://platform.openai.com/docs/models/gpt-4-turbo?snapshot=gpt-4-1106-vision-preview [accessed 2025-08-26]
34. Med-PaLM. Google. URL: https://sites.research.google/med-palm [accessed 2025-08-14]
35. MedLM: generative AI fine-tuned for the healthcare industry. Google. URL: https://cloud.google.com/blog/topics/healthcare-life-sciences/introducing-medlm-for-the-healthcare-industry [accessed 2025-08-14]
36. Gemini models. Google. URL: https://ai.google.dev/gemini-api/docs/models [accessed 2025-08-14]

37. Cohere's Command R+ Model. Cohere. URL: https://docs.cohere.com/docs/command-r-plus [accessed 2025-08-14]
38. GPT-4o. OpenAI. URL: https://platform.openai.com/docs/models/gpt-4o [accessed 2025-08-14]
39. Claude 3.5 Sonnet. Anthropic. URL: https://www.anthropic.com/news/claude-3-5-sonnet [accessed 2025-08-14]
40. Mistral Large 2. Mistral AI. URL: https://mistral.ai/news/mistral-large-2407 [accessed 2025-08-14]
41. Introducing Llama 3.1. Meta. URL: https://ai.meta.com/blog/meta-llama-3-1 [accessed 2025-08-14]
42. o3-mini. OpenAI. URL: https://platform.openai.com/docs/models/o3-mini [accessed 2025-08-14]
43. Claude 3.7 Sonnet and Claude Code. Anthropic. URL: https://www.anthropic.com/news/claude-3-7-sonnet [accessed 2025-08-14]
44. GPT-4.5 Preview. OpenAI. URL: https://platform.openai.com/docs/models/gpt-4.5-preview [accessed 2025-08-14]
45. The Llama 4 herd: the beginning of a new era of natively multimodal AI innovation. Meta. URL: https://ai.meta.com/blog/llama-4-multimodal-intelligence [accessed 2025-08-14]
46. GPT-4.1. OpenAI. URL: https://platform.openai.com/docs/models/gpt-4.1 [accessed 2025-08-14]
47. o3. OpenAI. URL: https://platform.openai.com/docs/models/o3 [accessed 2025-08-14]
48. o4-mini. OpenAI. URL: https://platform.openai.com/docs/models/o4-mini [accessed 2025-08-14]
49. Claude Sonnet 4. Anthropic. URL: https://www.anthropic.com/claude/sonnet [accessed 2025-08-14]
50. Claude Opus 4. Anthropic. URL: https://www.anthropic.com/claude/opus [accessed 2025-08-14]
51. ABIM laboratory test reference ranges January 2025. American Board of Internal Medicine. URL: https://www.abim.org/Media/bfijryql/laboratory-reference-ranges.pdf [accessed 2025-08-14]
52. BenchmarkingLLMs/Claude/Claude35/ClaudeEval.ipynb. GitHub. URL: https://github.com/rhazes-dev/BenchmarkingLLMs/blob/main/Claude/Claude35/ClaudeEval.ipynb [accessed 2025-08-14]
53. BenchmarkingLLMs/OpenAI/GPT4o/05-13/OGRUN_Temp0_05_13_GPT4o_Eval.ipynb. GitHub. URL: https://github.com/rhazes-dev/BenchmarkingLLMs/blob/main/OpenAI/GPT4o/05-13/OGRUN_Temp0_05_13_GPT4o_Eval.ipynb [accessed 2025-08-14]
54. BenchmarkingLLMs/OpenAI/GPT4o/05-13/TempdefPromptB/NewRun2_Tempdef_05_13_GPT4oProd_DiagAndEval.ipynb. GitHub. URL: https://github.com/rhazes-dev/BenchmarkingLLMs/blob/main/OpenAI/GPT4o/05-13/TempdefPromptB/NewRun2_Tempdef_05_13_GPT4oProd_DiagAndEval.ipynb [accessed 2025-08-14]
55. BenchmarkingLLMs/Gemini/Gemini2/TempdefGemini20ProdEval.ipynb. GitHub. URL: https://github.com/rhazes-dev/BenchmarkingLLMs/blob/main/Gemini/Gemini2/TempdefGemini20ProdEval.ipynb [accessed 2025-08-14]
56. BenchmarkingLLMs/Claude/Claude37/Temp0_Claude37Prod.ipynb. GitHub. URL: https://github.com/rhazes-dev/BenchmarkingLLMs/blob/main/Claude/Claude37/Temp0_Claude37Prod.ipynb [accessed 2025-08-14]
57. BenchmarkingLLMs/OpenAI/GPT4o/11-20/Tempdef_11_20_GPT4oProd_DiagAndEval.ipynb. GitHub. URL: https://github.com/rhazes-dev/BenchmarkingLLMs/blob/main/OpenAI/GPT4o/11-20/Tempdef_11_20_GPT4oProd_DiagAndEval.ipynb [accessed 2025-08-14]
58. BenchmarkingLLMs/OpenAI/GPT4o/11-20/Temp0_11_20_GPT4oProd_DiagAndEval.ipynb. GitHub. URL: https://github.com/rhazes-dev/BenchmarkingLLMs/blob/main/OpenAI/GPT4o/11-20/Temp0_11_20_GPT4oProd_DiagAndEval.ipynb [accessed 2025-08-14]
59. BenchmarkingLLMs/Claude/Claude37/Claude37Prod.ipynb. GitHub. URL: https://github.com/rhazes-dev/BenchmarkingLLMs/blob/main/Claude/Claude37/Claude37Prod.ipynb [accessed 2025-08-14]
60. BenchmarkingLLMs/OpenAI/GPT45/GPT45ProdDiagAndEval.ipynb. GitHub. URL: https://github.com/rhazes-dev/BenchmarkingLLMs/blob/main/OpenAI/GPT45/GPT45ProdDiagAndEval.ipynb [accessed 2025-08-14]
61. BenchmarkingLLMs/OpenAI/GPT45/Temp0_GPT45ProdDiagAndEval.ipynb. GitHub. URL: https://github.com/rhazes-dev/BenchmarkingLLMs/blob/main/OpenAI/GPT45/Temp0_GPT45ProdDiagAndEval.ipynb [accessed 2025-08-14]
62. BenchmarkingLLMs/OpenAI/GPT41/GPT41ProdDiagAndEval.ipynb. GitHub. URL: https://github.com/rhazes-dev/BenchmarkingLLMs/blob/main/OpenAI/GPT41/GPT41ProdDiagAndEval.ipynb [accessed 2025-08-14]
63. BenchmarkingLLMs/41eval/o4-mini/41eval_o4mini_ProdDiagAndEval.ipynb. GitHub. URL: https://github.com/rhazes-dev/BenchmarkingLLMs/blob/main/41eval/o4-mini/41eval_o4mini_ProdDiagAndEval.ipynb [accessed 2025-08-14]
64. BenchmarkingLLMs/41eval/GPT-4o-05-13/41eval_Tempdef_05_13_GPT4oProd_DiagAndEval.ipynb. GitHub. URL: https://github.com/rhazes-dev/BenchmarkingLLMs/blob/main/41eval/GPT-4o-05-13/41eval_Tempdef_05_13_GPT4oProd_DiagAndEval.ipynb [accessed 2025-08-14]
65. BenchmarkingLLMs/41eval/GPT-4o-05-13/Run1_41eval_Tempdef_05_13_GPT4oProd_DiagAndEval.ipynb. GitHub. URL: https://github.com/rhazes-dev/BenchmarkingLLMs/blob/main/41eval/GPT-4o-05-13/Run1_41eval_Tempdef_05_13_GPT4oProd_DiagAndEval.ipynb [accessed 2025-08-14]
66. BenchmarkingLLMs/41eval/GPT-4o-05-13/Run2_41eval_Tempdef_05_13_GPT4oProd_DiagAndEval.ipynb. GitHub. URL: https://github.com/rhazes-dev/BenchmarkingLLMs/blob/main/41eval/GPT-4o-05-13/Run2_41eval_Tempdef_05_13_GPT4oProd_DiagAndEval.ipynb [accessed 2025-08-14]
67. BenchmarkingLLMs/41eval/LLaMa4/41eval_LLaMa4ScoutProdEval.ipynb. GitHub. URL: https://github.com/rhazes-dev/BenchmarkingLLMs/blob/main/41eval/LLaMa4/41eval_LLaMa4ScoutProdEval.ipynb [accessed 2025-08-14]
68. BenchmarkingLLMs/41eval/Claude/41eval_Claude4Prod.ipynb. GitHub. URL: https://github.com/rhazes-dev/BenchmarkingLLMs/blob/main/41eval/Claude/41eval_Claude4Prod.ipynb [accessed 2025-08-14]

69. BenchmarkingLLMs/41eval/Claude/41eval_ClaudeOpus4Prod.ipynb. GitHub. URL: https://github.com/rhazes-dev/BenchmarkingLLMs/blob/main/41eval/Claude/41eval_ClaudeOpus4Prod.ipynb [accessed 2025-08-14]
70. BenchmarkingLLMs/41eval/o3/41eval_o3_ProdDiagAndEval.ipynb. GitHub. URL: https://github.com/rhazes-dev/BenchmarkingLLMs/blob/main/41eval/o3/41eval_o3_ProdDiagAndEval.ipynb [accessed 2025-08-14]
71. BenchmarkingLLMs/41eval/GPT-4.1/41eval_GPT41ProdDiagAndEval.ipynb. GitHub. URL: https://github.com/rhazes-dev/BenchmarkingLLMs/blob/main/41eval/GPT-4.1/41eval_GPT41ProdDiagAndEval.ipynb [accessed 2025-08-14]
72. BenchmarkingLLMs/41eval/Gemini2.5/41eval_Gemini25ProdEval.ipynb. GitHub. URL: https://github.com/rhazes-dev/BenchmarkingLLMs/blob/main/41eval/Gemini2.5/41eval_Gemini25ProdEval.ipynb [accessed 2025-08-14]
73. Shah-Mohammadi F, Finkelstein J. Accuracy evaluation of GPT-assisted differential diagnosis in emergency department. Diagnostics (Basel) 2024 Aug 15;14(16):1779. [doi: 10.3390/diagnostics14161779] [Medline: 39202267]
74. Ouyang L, Wu J, Jiang X. Training language models to follow instructions with human feedback. Presented at: 36th International Conference on Neural Information Processing Systems; Nov 28 to Dec 9, 2022; New Orleans, LA. [doi: 10.5555/3600270.3602281]
75. Gemini 2.5: pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. Google. URL: https://storage.googleapis.com/deepmind-media/gemini/gemini_v2_5_report.pdf [accessed 2025-08-14]
76. Maher J. English as an international language of medicine. Med Educ 1987 Jul;21(4):283-284. [doi: 10.1111/j.1365-2923.1987.tb00363.x] [Medline: 3626893]
77. Chimirri L, Caufield JH, Bridges Y, et al. Consistent performance of GPT-4o in rare disease diagnosis across nine languages and 4967 cases. medRxiv. Preprint posted online on Feb 28, 2025. [doi: 10.1101/2025.02.26.25322769] [Medline: 40061308]
78. Sarvarip/MIMIC-SQL. GitHub. URL: https://github.com/sarvarip/MIMIC-SQL [accessed 2025-08-14]
79. Rhazes-dev/benchmarkingllms. GitHub. URL: https://github.com/rhazes-dev/BenchmarkingLLMs [accessed 2025-08-14]

Abbreviations

AI: artificial intelligence
API: application programming interface
CCSR: Clinical Classifications Software Refined
ECG: electrocardiography
***ICD*:** *International Classification of Diseases*
***ICD-10*:** *International Classification of Diseases, 10th Revision*
LLM: large language model
MIMIC-IV: Medical Information Mart for Intensive Care-IV
MoE : Mixture of Experts
NEJM: *New England Journal of Medicine*
RAG: retrieval-augmented generation

Edited by A Schwartz; submitted 17.10.24; peer-reviewed by D Saderi, RS Gomaa Mahmoud, G Bender, OO Oladoyin, PH Ilegbusi, A Rahgozar, M Roy, M Machado, B Senst, CA Nkpoikanke Akpan, N Shaballout, S Sakilay, S Vohra, M Collier, M Olalekan Raimi, UK Chalwadi; revised version received 14.07.25; accepted 23.07.25; published 29.08.25.

Please cite as:
Sarvari P, Al-fagih Z
Rapidly Benchmarking Large Language Models for Diagnosing Comorbid Patients: Comparative Study Leveraging the LLM-as-a-Judge Method
JMIRx Med 2025;6:e67661
URL: https://xmed.jmir.org/2025/1/e67661
doi:10.2196/67661

Peer Review of "Safety and Efficacy of Chimeric Antigen Receptor T-Cell Therapy for Recurrent Glioblastoma: An Augmented Meta-Analysis of Phase 1 Clinical Trials (Preprint)"

Vanessa Fairhurst[1]; Randa Salah Gomaa Mahmoud[2]; Toba Olatoye[3]; Sylvester Sakilay

[1]PREreview, -, Portland, OR, United States

[2]Faculty of Human Medicine, Zagazig University, Zagazig, Egypt

[3]University of Ilorin, Ilorin, Nigeria

Related Article:

Companion article: https://www.medrxiv.org/content/10.1101/2024.10.23.24316015v1

(JMIRx Med 2025;6:e71293) doi:10.2196/71293

KEYWORDS

CAR T-cell therapy; cancer; glioblastoma; brain tumor; meta-analysis; chimeric antigen receptor

This is a peer-review report for the preprint "Safety and Efficacy of Chimeric Antigen Receptor T-cell Therapy for Recurrent Glioblastoma: An Augmented Meta-Analysis of Phase 1 Clinical Trials."

This review is the result of a virtual collaborative live review discussion organized and hosted by PREreview and JMIR Publications on Dec 12, 2024. The discussion was joined by 11 people: 3 facilitators, 1 member of the JMIR Publications team, and 7 live review participants including 3 who agreed to be named but did not assist in compiling the final review: Eudora Nwanaforo, Kelechi Elechi, and Murtala Haruna Bawa. The authors of this review have dedicated additional asynchronous time over the course of 2 weeks to help compose this final report using the notes from the live review. We thank all participants who contributed to the discussion and made it possible for us to provide feedback on this preprint.

Summary

The study [1] was designed to address the limitations of previous studies and evaluate the safety and efficacy of chimeric antigen receptor (CAR) T-cell therapy for recurrent glioblastoma. The results of this study are predictive rather than confirmatory. CAR T-cell therapy for glioblastoma was not predicted to significantly improve survival or achieve substantial complete responses. Stable disease rates were modest, while disease progression was notable. Adverse events, especially CAR T-cell therapy–related encephalopathy, raise safety concerns. Overall survival was 6.49 months in patients receiving CAR T-cell therapy after augmented analysis, and only 80% of patients exhibited this outcome. It was not statistically different from the median overall survival observed in patients with recurrent glioblastoma undergoing standard treatment, thereby indicating that CAR T-cell therapy, in its current form, does not offer substantially improved survival compared to standard treatments. Further trials and refinements are needed to enhance CAR T-cell therapy's effectiveness and safety in glioblastoma treatment. An interesting fact is that a novel statistical technique (augmented meta-analyses) was used in this study. It was a combination of a cross-sectional (quantitative) and augmented meta-analysis (qualitative).

List of Major Concerns and Feedback

Methods

Augmented Meta-Analysis

- This section is limited in its description of the methodology used in the study. It would be helpful to include more information on the machine learning model or language model used to generate the extra cases.
- The title and aim specify that the study focuses on recurrent glioblastoma, but this specificity is not reflected in the inclusion criteria. It would be helpful to adjust the inclusion criteria to explicitly state that the study is targeting patients with recurrent glioblastoma. This will align the methodology with the aim as stated.
- The inclusion criteria do not specify that patients are in phase 1 clinical trials, where safety is a primary focus. Clearly state in the inclusion criteria that patients are part of phase 1 clinical trials. This will provide context for the study's focus on safety.
- There is no reference to the earlier use of augmented meta-analysis in cancer or medical research, nor is it explicitly stated if this is a new application. If augmented meta-analysis has been previously applied, cite relevant references. If this is its first application, explicitly state so and highlight its novelty.

Results

Literature Review and Risk of Bias Assessment Section

- It would be helpful to add the details of Figure 1 and Table 1 that explain the details of the cause of exclusion, the

results of the Newcastle Ottawa Scale, which study reached the high-quality level, etc.

Discussion

- It is important to add a comparison between the mean overall survival for patients with glioblastoma who underwent CAR T-cell therapy and the median overall survival observed in patients receiving the standard protocol for recurrent glioblastoma treatment to the Results section, as this comparison is mentioned in the first paragraph of the Discussion section.

Reproducibility of the Study

- The data presented in the study are beneficial for reproducibility except for the augmented meta-analysis, which is hindered by the lack of clear documentation on the large language model settings.
- The details of the augmented meta-analysis are not available. Provide access to the source code or methodological details for augmented meta-analysis, either as supplementary material or a public repository link. Transparency will strengthen the study's reproducibility.

List of Minor Concerns and Feedback

Concerns With Techniques/Analyses

- Abbreviations like "IL-13Ralpha-2," "EGFRvIII," "HER2," and "HephA2" are not identified in the Included Study Characteristics section. Expand the abbreviations and provide their full names (eg, "Interleukin-13 Receptor Subunit Alpha-2") when first mentioned. This ensures clarity for readers not familiar with the terms.
- The last line of the large language model statement on page 16 does not explain how augmented meta-analysis was applied. Elaborate on how augmented meta-analysis was applied, especially in terms of methodology and integration with the study data.

Figures and Tables

- The screening section in Figure 1 is missing a rectangle to indicate the exclusion of 300 records. Update it using the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flowchart to include a rectangle that details the 300 excluded records and ensures the causes of exclusion are clearly stated.
- The reasons for exclusion are not detailed in the PRISMA flowchart. Follow PRISMA guidelines to specify the causes of exclusion, such as duplicates, irrelevance, or incomplete data, within the flowchart.
- Comments following Figure 1 are not in line with its instructions. Restructure the comments to follow the instructions and present the details of the research study accordingly.

Additional Comments

- No reference is provided for the trim-and-fill method mentioned in the augmented meta-analysis of overall survival (page 10). Cite a relevant source, such as [2] or another appropriate reference.
- The Cochrane Handbook (Part 2, Chapter 9) should be referenced in the Statistical Analysis section and its numbered reference cited in the text.
- References in the third paragraph of the Introduction mix meta-analyses and clinical trials without clear distinction. Rearrange and clarify the references while ensuring that references to meta-analyses and clinical trials are grouped and contextualized appropriately to avoid confusion.
- Repetition of the sentence "Egger's test for publication bias could not be performed since the number of included studies in this outcome was less than ten" could be avoided by mentioning it once in the Methods section as the total number of the included studies is 8.
- In addition, the repetition of the sentence "The wide range of the 95% confidence interval was suggestive of data sparsity, so augmented meta-analysis was indicated before making conclusions" could be avoided by mentioning it once in the Augmented Meta-Analysis section of the Methods.

Acknowledgments

PREreview and JMIR Publications thank the authors of the preprint for posting their work openly for feedback. We also thank all participants of the live review call for their time and for engaging in the lively discussion that generated this review.

Conflicts of Interest

VF was a facilitator of this call and one of the organizers. No other competing interests were declared by the reviewers.

References

1. Azzam AY, Morsy MM, Azab MA, et al. Safety and efficacy of chimeric antigen receptor T-cell therapy for recurrent glioblastoma: an augmented meta-analysis of phase 1 clinical trials. medRxiv. Preprint posted online on Oct 24, 2024. [doi: 10.1101/2024.10.23.24316015]
2. Shi L, Lin L. The trim-and-fill method for publication bias: practical guidelines and recommendations based on a large database of meta-analyses. Medicine (Baltimore) 2019 Jun;98(23):e15987. [doi: 10.1097/MD.0000000000015987] [Medline: 31169736]

Abbreviations

CAR: chimeric antigen receptor
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses

Edited by A Schwartz; submitted 14.01.25; this is a non–peer-reviewed article;accepted 14.01.25; published 24.01.25.

Please cite as:
Fairhurst V, Mahmoud RSG, Olatoye T, Sakilay S
Peer Review of "Safety and Efficacy of Chimeric Antigen Receptor T-Cell Therapy for Recurrent Glioblastoma: An Augmented Meta-Analysis of Phase 1 Clinical Trials (Preprint)"
JMIRx Med 2025;6:e71293
URL: https://xmed.jmir.org/2025/1/e71293
doi:10.2196/71293

Peer Review of “State Anxiety Biomarker Discovery: Electrooculography and Electrodermal Activity in Stress Monitoring (Preprint)”

Daniela Saderi[1]; Shailee Rasania[2]; Toba Olatoye[3]; Simon Muhindi Savai; Randa Salah Gomaa Mahmoud[4]; Vasco Medeiros; Mitchell Collier

[1]PREreview, Portland, OR, United States

[2]Inflammatix, Inc, King of Prussia, PA, United States

[3]Kwara State Teaching Service Commission, Ilorin, Nigeria

[4]Zagazig University, Zagazig, Egypt

Related Article:

Companion article: https://arxiv.org/abs/2411.17935v1

(JMIRx Med 2025;6:e72093) doi:10.2196/72093

KEYWORDS

stress; biomarker discovery; EOG; EEG; medical informatics; electrooculography; electroencephalography

This is a peer-review report for the preprint “State Anxiety Biomarker Discovery: Electrooculography and Electrodermal Activity in Stress Monitoring.”

This review is the result of a virtual collaborative live review discussion organized and hosted by PREreview and JMIR Publications on January 16, 2025. The discussion was joined by 16 people: 2 facilitators, 1 member of the JMIR Publications team, and 13 live review participants, including 3 who agreed to be named but have not contributed to composing this review into its final form: Uday Kumar Chalwadi, Killivalavan Solai, and Prasakthi Venkatesan. The authors of this review have dedicated additional asynchronous time over the course of 2 weeks to help compose this final report using the notes from the live review. We thank all participants who contributed to the discussion and made it possible for us to provide feedback on this preprint.

Summary

Anxiety, particularly state anxiety (s-anxiety), is increasingly recognized as a health concern linked to mental and physical issues, including adverse cardiovascular and long-term health outcomes. This study [1] leverages noninvasive wearable technology to identify interpretable biomarkers resulting from s-anxiety using electrooculography (EOG) and electrodermal activity (EDA). Two datasets were developed: BLINKEO, focusing on blink-related EOG features, and EMOCOLD, analyzing EOG and EDA responses during a cold pressor test. The authors then used both datasets and applied statistical analysis (eg, F_1-scoring, Shapley Additive Explanations [SHAP] analysis) to identify biomarkers of anxiety. Results revealed that using EOG data (blink duration, peak height, and opening integral) in tandem with EDA data (mean signal, permutation, entropy, and Hjorth activity) led to the identification of novel biomarkers that reveal nuanced emotional and stress responses. Moreover, it was found that SHAP analysis can more accurately determine which features are relevant to enhancing model performance. The findings highlight the potential of combining EOG and EDA biomarker data to create robust real-time models for anxiety detection. Combinations of physiological features (as sets) were more effective as measures of stress response than individual features alone. This research underscores the transformative role of noninvasive wearable technology in personalized mental health monitoring and intervention strategies.

List of Major Concerns and Feedback

Concerns With Methods

- It would be helpful to document the name of the device and manufacturer used to record the EOG. This would help other researchers who may want to reproduce the results.
- Similarly, it would be helpful to add additional details about the cold pressor test methods. For example, was a commercially available circulating water bath used to maintain a constant water temperature? Was the temperature of the subject’s hand monitored? The details of the cold stressor test (the water temperature, the period of immersion, and the cutoff point) should be added for the sake of clarity, transparency, and reproducibility. Past studies using these metrics should also be referenced for details (eg, [2]). These methodological details may also be added in the form of a figure to add clarity to the experimental setup.
- To better understand the individual response to the cold challenge before participating in the actual experiment, it is advised that the manuscript states what type of participant testing was or was not adopted in the cold pressor testing experiment. For example, what were the tolerance times?

Were there any gender differences? If any pretesting data were collected, analyzing them and presenting them as results would add clarity to the results.
- It is unclear if the 65 repeating blinking trials and the 19 no-blinking trials were collected from the same individual or from different individuals. Please clarify.
- No signal voltage/electrical records for EDA were found in the manuscript. Is this intentional? Please consider adding this information.
- It would be important to add details of ordinal variables present in the Positive and Negative Affect Schedule and the State-Trait Anxiety Inventory (STAI-State), and clearly state their function and use in Supplementary Table 2.

Concerns With Analysis

- F_1-scores that were mentioned in the text (87.34% and 79.99%) are not present within the figures. Moreover, an F_1-score is an integer value from 0 to 1, taking precision and recall into account, and is not often expressed as a percentage.
- Figure 1c has two separate graphs; it should be captioned as 1c and 1d. What do both these graphs portray? The second graph for 1c is missing titles for the x- and y-axes—the current assumption is that they are the same as the first graph.
- Table 1 lacks a legend and is shown as panel a of Table 2. Please check how the tables are referenced in the text to make sure they reference the right one.
- The captions of the figures should have statistical information when relevant. For example, in Figure 3, the caption should include a description of what data were plotted and the meaning of the graph. Presumably plotting medians, quartiles, and SDs? Also, please report n values.

Concerns With Ethics

- It is not clear what the ethical statement at the end of the manuscript, which states that the study was exempt from review board approval, means. That statement should be revised for clarification. In addition, details regarding whether or not institutional review board approval was obtained, whether the study involved consenting participants and used humans, how the data were collected and used, how the data were handled to protect the privacy of study participants, and any other ethical procedures that were followed to protect subjects from any harm due to participation in the study should be added.

List of Minor Concerns and Feedback

Minor Concerns With Methods

- Please document whether the data were taken from each subject only once or whether data were obtained several times from a subject.
- Referring to the line “To focus on blink-like events, we applied criteria based on established blink characteristics,” the criteria used to establish blink characteristics should be cited, if not already given.
- SHAP analysis was performed on combinations of 5 features. Please clarify on what basis these 5 features were chosen (out of 15 of EDG and 33 of EOG).

Minor Concerns With Analysis and Presentation

- Page 10, Electrooculography (EOG) Signal Segmentation section: the authors mentioned that they extracted 33 features; however, Supplementary 4 mentioned 35 feature definitions. Please revise and correct.
- In Figure 3, please put “STAI-State survey score” on the y-axis for clarification rather than just “Scores.” In addition to box and whiskers plots, adding column graphs for positive affectivity, negative affectivity, and s-anxiety might be beneficial to more clearly express the SD present within the data.
- It would be beneficial to graphically display the F_1-scores that were collected across the study.
- The figures are quite small, which makes readability a little difficult. Please make the text larger to improve readability and accessibility.
- The Figure 1a description states, “The red dotted lines indicate the center of the peak…,” but these appear to be gray.

Suggestions

- Consider the inclusion of a Limitations section in this manuscript to better discuss potential limitations due to the skewness in male and female participants, data curation, applied methodologies, and other limitations of the study.
- A figure showing the trial structure would be very useful to understand how the data were collected.

References

- In the third paragraph of the Introduction, adding a reference to other techniques used to provoke anxiety, including the reduced EDA response in depressed patients, and the conflicting studies could be helpful to the readers.
- In the Introduction, fourth paragraph, the reference “Schachter and Singer” is not present in the References. Is this the wrong reference, or it just needs to be added to the list?
- In the Introduction, third page, third paragraph, it is advised to add references to document the reduced EDA response in depressed patients and the conflicting studies.
- In the Methods, please cite sources for the Butterworth filter (page 5), the Savitzky-Golay filter (page 5), and all other analyses.
- Reference 2: Include full citation with a link.
- Reference 3: It is advised to correct the article name to “APA 2023 Stress in America Topline Data.”
- Reference 4: The correct citation should be “Kazanskiy NL., Khonina S.N., Butt M.A. A review on flexible wearables—Recent developments in non-invasive continuous health monitoring. Sens. Actuators A Phys. 2024;366:114993. doi: 10.1016/j.sna.2023.114993.”
- Reference 10: The correct citation should be: “Electrooculogram Analysis and Development of a System for Defining Stages of Drowsiness Master's Thesis Project

in Biomedical Engineering, Linköping University, Dept. Biomedical Engineering, LiU-IMT-EX-351 Linköping 2003. Available: https://www.diva-portal.org/smash/get/diva2:673960/FULLTEXT01.pdf Test"

- Reference 19: The correct citation should be "Anxiety Detection Using Multimodal Physiological Sensing, 2021 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI), Athens, Greece, 2021, pp. 1-4, doi: 10.1109/BHI50953.2021.9508589."
- Reference 23: Revising this citation is advised as searching on the internet shows error 404. The requested URL was not found on this server. Moreover, this is not a proper citation—give the edition number of the book (there are at least 5 editions) and publication year, as well as the page number of the cited data point about typical blink elapsed time.
- Reference 27: The correct citation should be "Hassanein, A.M.D.E., Mohamed, A.G.M.A. & Abdullah, M.A.H.M. Classifying blinking and winking EOG signals using statistical analysis and LSTM algorithm. Journal of Electrical Systems and Inf Technol 10, 44 (2023). https://doi.org/10.1186/s43067-023-00112-2."
- In general, citations need to be reviewed and added with consistency throughout the manuscript.

Acknowledgments

PREreview and JMIR Publications thank the authors of the preprint for posting their work openly for feedback. We also thank all participants of the live review call for their time and for engaging in the lively discussion that generated this review.

Conflicts of Interest

DS was a facilitator of this call and one of the organizers. No other competing interests were declared by the reviewers.

References

1. Dao J, Liu R, Solomon S, Solomon S. State anxiety biomarker discovery: electrooculography and electrodermal activity in stress monitoring. arXiv. Preprint posted online on Nov 26, 2024. [doi: 10.48550/arXiv.2411.17935]
2. Mitchell LA, MacDonald RAR, Brodie EE. Temperature and the cold pressor test. J Pain 2004 May;5(4):233-237. [doi: 10.1016/j.jpain.2004.03.004] [Medline: 15162346]

Abbreviations

EDA: electrodermal activity
EOG: electrooculography
s-anxiety: state anxiety
SHAP: Shapley Additive Explanations
STAI: State-Trait Anxiety Inventory

Edited by A Schwartz; submitted 03.02.25; this is a non–peer-reviewed article;accepted 03.02.25; published 03.03.25.

Please cite as:
Saderi D, Rasania S, Olatoye T, Savai SM, Mahmoud RSG, Medeiros V, Collier M
Peer Review of "State Anxiety Biomarker Discovery: Electrooculography and Electrodermal Activity in Stress Monitoring (Preprint)"
JMIRx Med 2025;6:e72093
URL: https://xmed.jmir.org/2025/1/e72093
doi:10.2196/72093

Peer Review of "Large Language Models for Pediatric Differential Diagnoses in Rural Health Care: Multicenter Retrospective Cohort Study Comparing GPT-3 With Pediatrician Performance"

Daniela Saderi[1]; Goktug Bender[2]; Toba Olatoye[3]; Arya Rahgozar[4], PhD; Uday Kumar Chalwadi[5]; Eudora Nwanaforo[6]; Paul Hassan Ilegbusi[7]; Sylvester Sakilay; Mitchell Collier

[1]PREreview, Portland, OR, United States
[2]McGill University, Montreal, ON, Canada
[3]University of Ilorin, Ilorin, Nigeria
[4]University of Ottawa, Ottawa, ON, Canada
[5]LSUHS, Shreveport, LA, United States
[6]Federal University of Technology, Owerri, Nigeria
[7]Ondo State College of Health Technology, Akure, Nigeria

Related Articles:

Companion article: https://www.medrxiv.org/content/10.1101/2024.08.09.24311777v1

Companion article: https://med.jmirx.org/2025/1/e73258

Companion article: https://med.jmirx.org/2025/1/e65263

(JMIRx Med 2025;6:e73264) doi:10.2196/73264

KEYWORDS

natural language processing; NLP; machine learning; ML; artificial intelligence; language model; large language model; LLM; generative pretrained transformer; GPT; pediatrics

This is the peer-review report for "Large Language Models for Pediatric Differential Diagnoses in Rural Health Care: Multicenter Retrospective Cohort Study Comparing GPT-3 With Pediatrician Performance."

This review is the result of a virtual collaborative live review organized and hosted by PREreview and JMIR Publications on October 25, 2024. The discussion was joined by 21 people: 2 facilitators, 1 member of the JMIR Publications team, and 18 live review participants, including 3 who agreed to be named here but did not contribute to writing this review: Nour Shaballout, Randa Salah Gomaa Mahmoud, and Samaila Jackson Yaga. The authors of this review have dedicated additional asynchronous time over the course of 2 weeks to help compose this final report using the notes from the live review. We thank all participants who contributed to the discussion and made it possible for us to provide feedback on this preprint.

Summary

The study [1] seeks to determine how accurately and reliably a fine-tuned GPT-3 model can assist with differential diagnosis in pediatric cases within rural health care environments. Specifically, it examines whether the artificial intelligence (AI) model can match or approach the diagnostic accuracy of human physicians. By evaluating the model's diagnostic performance, the research aims to explore AI's potential to improve pediatric health care quality, reduce misdiagnosis, and support providers in underserved regions where accurate, timely diagnosis is critical for patient outcomes.

To address the research questions, the authors conducted a retrospective study using data from 500 pediatric cases from a multicenter rural pediatric health care organization in Central Louisiana, United States. The GPT-3 model was trained on 70% of the data, including symptoms and physician-provided differential diagnoses, and tested on the remaining 30%, achieving an accuracy of 87%, with sensitivity at 85% and specificity at 90%. These results were statistically comparable to human physicians, who had an accuracy of 91%. The findings suggest that AI can support clinical decision-making in pediatric care, especially in resource-constrained environments where access to specialists is limited.

The research addresses critical gaps in pediatric care by exploring AI's potential to support clinical decision-making, particularly in resource-limited settings. It presents this with methodological details that enhance reproducibility and offer insights into AI applications in health care. The authors' transparency about limitations reflects research integrity, establishing a strong base for future studies. Furthermore, the focus on integrating AI into clinical workflows shows an understanding of practical challenges and underscores opportunities for advancing health care delivery through

technology. However, the study presents some notable weaknesses, including a lack of assessment of patient outcomes and insufficient clarity in its methodology, indicating areas for future research and improvement. Below, we list specific concerns and recommendations on how to address them.

List of Major Concerns and Feedback

Concerns With Techniques and Analyses

- Model choice: It is unclear why a specific generative AI model (ie, GPT-3, DaVinci version) was chosen for this study. Was the GPT-3 model (DaVinci version) selected due to its extensive use in medical AI research, or was it chosen to facilitate comparison with previous studies? A statement explaining the choice of the AI model would significantly improve the reader's understanding of the study's context and its relationship to previous research.
- Normality test: The study does not address whether data normality was assessed before statistical analysis. Determining the distribution of the data is key to selecting the appropriate statistical test to analyze such data. The Kolmogorov-Smirnov test could aid in understanding data distribution, specifically testing for normality. If the data is not found to meet normality criteria, nonparametric methods should be applied. Including a data normality assessment and explaining the choice of a particular statistical test would significantly strengthen the reliability of the study.
- Evaluation metrics: The study primarily uses specificity and sensitivity for evaluating large language model–generated responses, which may not capture the full quality of the outputs. Incorporating natural language processing metrics such as Recall-Oriented Understudy for Gisting Evaluation (ROUGE) and bilingual evaluation understudy (BLEU) can help assess the quality of generated responses more comprehensively. ROUGE measures the correspondence between the automatically generated response versus that of the human and what was expected. There are also issues associated with large language model generations of responses such as hallucination and the lack of attribution. Please specify or comment on how those and other issues were measured.
- Power analysis assumptions: The assumptions underlying the power analysis are unclear, particularly regarding how specific diagnoses affect this analysis. It is advised to elaborate on the power analysis methodology, including the rationale behind sample size choices and their implications for diagnosis variability.
- Sample size and generalizability: The sample size of 500 encounters may not adequately represent the broader pediatric population, particularly in diverse settings. Furthermore, using data from a single health care organization limits the applicability of findings to other settings. These limitations should be discussed, particularly how the validity of the results might change when it is tested with data from other health care centers. If possible, authors should mention and cite studies that reported on this effect. Additionally, future studies should consider expanding the sample size through multicenter collaborations or including data from patients with more diverse demographics to validate results across different health care environments thereby enhancing generalizability.

Details for Reproducibility of the Study

- Software and tools documentation: The authors describe using both Python (with scikit-learn) and IBM SPSS Statistics, but it is unclear what the software's sources are. Specifying sources for Python and scikit-learn (eg, "Python 3.8 [Python Software Foundation, Delaware, USA]") and clarifying the respective roles of Python and SPSS in the analyses would enhance transparency and allow for the reproducibility of the study.
- Detailed group descriptions: The demographics, specifically age group cases, are underspecified, limiting the reader's understanding of the study sample. Adding a table or descriptive text detailing subgroup demographics, including age and case counts would improve the study's interpretability and allow readers to better contextualize findings.
- Cross-validation across organizations: The model's reproducibility across various health care settings is not demonstrated. Evidence shows models often underperform with data from different sources. Including cross-organization validation and clearly acknowledging this limitation in the Discussion by citing relevant studies would enhance robustness. Furthermore, addressing this limitation in future work could pave the way for broader adoption and application of the model.
- Data and model specifics for replicability: The study would benefit from more thorough descriptions of dataset characteristics, fine-tuning model parameters, and preprocessing methods. For validation, consider adding multicenter dataset details. Adding this information would enable other researchers to replicate and build upon the study's findings, thereby enhancing its scientific contribution.
- Diagnostic exclusion or inclusion clarification: The preprocessing section does not clarify if physician diagnostics were included or excluded, leading to potential confusion for readers and impacting reproducibility. It would be helpful to know whether physician diagnostics were included in training and why. Clarifying this aspect would help standardize study replication and improve the study's transparency.

Figures and Tables

- Figure 1 is mentioned but not included in the article, which affects comprehension of the study design and findings. Please include Figure 1 or provide an alternative reference to explain the content of the missing figure. Figures are helpful for readers to quickly grasp complex methodologies and findings.

Ethics

- Data privacy: It is unclear whether a private or public version of GPT-3 was used, and if the latter, this raises potential Health Insurance Portability and Accountability Act (HIPAA) concerns. As was already pointed out above,

it is recommended that the version of GPT-3 used is specified, with additional clarification regarding data privacy practices if a public model was used. The addition of HIPAA considerations will enhance readers' confidence in the study's privacy protocols.
- Discussion of diagnostic risk: The discussion would benefit from a deeper exploration of diagnostic risks associated with the use of AI in health care and clinical decision-making settings. One example is the potential of AI models to perpetuate and affirm existing human biases thereby further exacerbating health disparities (one relevant citation could be Mittermaier M, Raza MM, Kvedar JC. Bias in AI-based models for medical applications: challenges and mitigation strategies. *NPJ Digit Med.* Jun 14, 2023;6(1):113 [doi: 10.1038/s41746-023-00858-z] [Medline: 37311802]). The study also raises important social considerations, such as respecting human agency, particularly for vulnerable populations. Addressing parental concerns about deferring decision-making to AI is crucial, as is ensuring a socially attuned approach to building trust and understanding.
- Lack of clarity on potential implementation in rural health care settings: The study could be strengthened by detailing how the AI model might be implemented in rural health care settings, including the specific challenges involved. Key considerations include the need for sufficient infrastructure (eg, electricity, internet) and the necessity of training health care providers unfamiliar with AI tools. Additionally, discussing both the potential impact (eg, improved diagnostic efficiency) and limitations (eg, handling incomplete data or overreliance on AI) would provide a more comprehensive road map for deployment in rural environments.

List of Minor Concerns and Feedback

- Data distribution gaps: No comparison of racial identity distribution between training and testing sets. Please consider adding a table or section on these demographic comparisons to ensure representation across subgroups.
- Data description and context: It would be helpful to know more information regarding how physicians were selected and their specific roles in the study.
- Departmental affiliations: Authors' affiliations lack specific department details, which limits transparency. Include departmental affiliations for authors to increase transparency and traceability. Adding departmental affiliations will provide context on the authors' expertise and institutional support.
- Funding transparency: The funding statement does not clearly specify whether the study was internally or externally funded. Explicitly state funding details, clarifying internal/external sources as applicable. Clear funding information will enhance transparency and address potential conflicts of interest.
- Approval number: While an ethical approval statement is present, it lacks the approval number, which is critical for ethical transparency. Please include the ethics approval number/code to ensure proper documentation and strengthen the study's validity and trustworthiness.
- Inconsistent data collection dates between the abstract and data collection section (lines 19 and 82)
- Missing figure (line 104).
- Need for more descriptive statistics (mean, median, quartiles, SD).
- Data distribution: Lack of comparison for racial/Hispanic identity distribution between training and testing sets. There's insufficient detail on age subgroup distribution.
- Clarification needed: The authors need to provide a deeper discussion of the power analysis methodology.
- The authors assessed that the distribution of age, gender, and chief complaints was similar between the training and testing sets. Suggest this to be cited to Table 5.
- Table 1: The abbreviations in the formula column should be identified in the table legend as "(FN: False Negative; FP: False Positive; TN: True Negative; TP: True Positive) (m)+1."
- Please clarify why GPT-3.5 or GPT-4 (instead of GPT-3) was not used despite being available at the time of the study.
- Line 103 states physicians were instructed to generate differential diagnoses. I thought this was obtained retrospectively. Please clarify.
- Line 152: Table 4 should be corrected to Table 3.
- Line 154: Table 5 should be corrected to Table 4.
- Line 200: Typo "may limit the of the finding."

Concluding Remarks

We thank the authors of the preprint for posting their work openly for feedback. We also thank all participants of the live review call for their time and for engaging in the lively discussion that generated this review.

Conflicts of Interest

DS contributed to writing this review and was a facilitator of this call and one of the organizers. No other competing interests were declared by other reviewers who participated in discussing the preprint during the live review.

Reference

1. Mansoor M, Ibrahim AF, Grindem D, Baig A. Large language models for pediatric differential diagnoses in rural health care: multicenter retrospective cohort study comparing GPT-3 with pediatrician performance. JMIRx Med 2025;6:e65263. [doi: 10.2196/65263]

Abbreviations

AI: artificial intelligence
BLEU: bilingual evaluation understudy
HIPAA: Health Insurance Portability and Accountability Act
ROUGE: Recall-Oriented Understudy for Gisting Evaluation

Edited by A Schwartz; submitted 28.02.25; this is a non–peer-reviewed article;accepted 28.02.25; published 19.03.25.

Please cite as:
Saderi D, Bender G, Olatoye T, Rahgozar A, Chalwadi UK, Nwanaforo E, Ilegbusi PH, Sakilay S, Collier M
Peer Review of "Large Language Models for Pediatric Differential Diagnoses in Rural Health Care: Multicenter Retrospective Cohort Study Comparing GPT-3 With Pediatrician Performance"
JMIRx Med 2025;6:e73264
URL: https://xmed.jmir.org/2025/1/e73264
doi:10.2196/73264

Peer Review of "The Order in Speech Disorder: A Scoping Review of State of the Art Machine Learning Methods for Clinical Speech Classification (Preprint)"

Vanessa Fairhurst[1]; Sylvester Sakilay; Randa Salah Gomaa Mahmoud[2]; Shailee Rasania; J Moonga[3]; Toba Isaac Olatoye[4]; Rameshwari Prasad[5]; Prasakthi Venkatesan; Vasco Medeiros[3]; Uday Kumar Chalwadi[6]

[1]PREreview, Portland, OR, United States
[2]Zagazig University, Zagazig, Egypt
[3]King's College London, London, United Kingdom
[4]Kwara State Teaching Service Commission, Ilorin, Nigeria
[5]Shelby County Health Department, Memphis, TN, United States
[6]Louisiana State University Health Sciences Center Shreveport, Shreveport, LA, United States

Related Article:

Companion article: https://arxiv.org/abs/2503.04802v1

(JMIRx Med 2025;6:e76836) doi:10.2196/76836

KEYWORDS

scoping review; machine learning; speech patterns; diagnosis; speech disorders; mental disorders; neurological disorders

This is the peer-review report for the preprint "The Order in Speech Disorder: A Scoping Review of State of the Art Machine Learning Methods for Clinical Speech Classification."

This review is the result of a virtual collaborative live review discussion organized and hosted by PREreview and JMIR Publications on April 10, 2025. The discussion was joined by 29 people: 3 facilitators from the PREreview team, 1 member of the JMIR Publications team, and 25 live review participants, 4 of whom joined as listeners and did not contribute to the review. The authors of this review have dedicated additional asynchronous time after the call over the course of 2 weeks to help compose this final report using the notes from the live review. We thank all participants who contributed to the discussion and made it possible for us to provide feedback on this preprint.

Summary

Speech is a cornerstone of human communication, intricately connected to our cognitive, neurological, and psychological processes. Speech patterns have emerged as potential diagnostic markers for conditions with varying etiologies. This scoping review [1] elucidates how machine learning (ML) can utilize speech patterns as noninvasive diagnostic biomarkers for neurological, laryngeal, and mental health etiologies. Based on specific inclusion and exclusion criteria that involved a wide spectrum of conditions, ranging from voice pathologies to mental and neurological disorders, the 564 articles compiled in this investigation were condensed to 91. Methods of speech classification were then assessed between 0 - 10 based on the diagnostic accuracy of different ML models. High accuracies were reported for Parkinson disease, laryngeal disorders, and dysarthria, whereas disorders like depression, schizophrenia, mild cognitive impairment, and Alzheimer disease (AD) showed promise yet were less consistent. This review emphasizes the need for speech analysis in conditions like obsessive-compulsive disorder and autism, where graded clinical diagnoses are less robust, relative to other disorders. Key strengths of the preprint include its comprehensive coverage of disorders and the current relevance of the literature (post 2016). However, noted limitations include a lack of cross-linguistic model generalizations, a limited coverage of pediatric populations, and sociocultural variations in speech. Despite some ambiguity present in the methodologies, the paper effectively bridges the fields of speech science, artificial intelligence (AI), and clinical diagnostics. Moreover, it highlights the transformative potential of ML in developing personalized scalable diagnostic models while also considering ethical implications, clinical acceptance, and real-world applications.

List of Major Concerns and Feedback

With "major concerns," we refer to concerns that the reviewers believe should be prioritized in being addressed in order to ensure the soundness of the study.

Below, we summarize major concerns raised by the live review participants, and whenever possible, we offer suggestions on how to address them.

1. A lack of model validation: More clarity should be provided to highlight the distinction between disease state/features and symptoms. For example, neurodegenerative diseases such as AD and Huntington disease have features similar to neuropsychiatric diseases—schizophrenia, depression, etc. While the symptoms and manifestations can overlap,

they are not the same thing; they differ in etiology and characteristics. The failure to delineate those characteristics weakens the study's overarching question and rationale from the start.

2. A scoping review is meant to provide a wide scope of the literature to map out data and synthesize findings for interpretation and appraisal. There is a major weakness in the findings presented in the tables. At present, the evidence provided does not sufficiently reflect the body of empirical evidence that is available in neurodegeneration, linguistics, and ML methods to achieve the goals in the study aims/objectives. To increase the strength of the analysis and improve the data disseminated in the tables, one option could be to combine the similarities in findings in each table. This task can also improve the presentation of the data in each table.
3. It is not clear why the search is restricted to the PubMed application programming interface and does not include other platforms such as MEDLINE (OVID), Embase (Elsevier), PsycINFO (OVID), CINAHL, Google Scholar, and Web of Science.
4. The methods and results should be reported in accordance with scoping review guidelines, such as PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews) [2].
5. The keywords identified to search in databases should be mentioned (it could be added as a supplementary file).
6. The time range of the search was not mentioned.
7. There is a lack of clarity between the neurodegenerative diseases and neuropsychiatric diseases; for example, AD and schizophrenia should be distinguished since AD progresses at various stages that do not necessarily resemble the features of schizophrenia.
8. The dataset size and ratio of healthy controls versus patients are important factors that are necessary to mention in Tables 1-3.
9. Clinical relevance: There is a need to review the profile/demographics of cohorts and groups of participants in the selected studies. This would help to demonstrate the time-course of disease/condition in their application to ML and the nature of the pool of data extracted in the analytical phase of the study (ie, data synthesis and interpretation). That is critical information that could be obtained in the data extraction stage (per PRISMA guidelines). By establishing the clinical relevance here, the paper can better argue how ML methods can help clinical speech classification in neurological and psychiatric diseases for diagnostic purposes.
10. In the inclusion criteria, articles published in English were mentioned, but non-English articles were also included in the study. An explanation for the inclusion of non-English articles was not provided by the authors. Additionally, the study deliberately focused on speech parameters, excluding the analysis of language content, which could provide a more holistic understanding of communicative aspects related to health conditions. Mentioned in 4.6.
11. False negatives: In evaluations, speech can appear healthy even if an individual has a serious health condition, making false negatives an important consideration. Speech-based diagnostics should be an addition to other diagnostic methods, not a stand-alone solution. Authors mentioned this in 4.7.3 as a limitation, but no such attempt was observed in the inclusion of related literature.
12. The authors effectively address key issues such as patient data privacy, informed consent, General Data Protection Regulation (GDPR) compliance, and clinical deployment risks associated with AI-driven speech diagnostics. The inclusion of synthetic speech data as a means to mitigate privacy concerns is a noteworthy strength. To enhance this section, we recommend incorporating specific frameworks or strategies—such as data anonymization, algorithmic transparency, and regulatory guidance—to provide a more robust and actionable ethical foundation for clinical implementation. Ethical considerations, especially around AI deployment, patient data privacy, and consent, should be discussed in more detail.
13. The manuscript provides valuable insights but would benefit from a more comprehensive discussion of its limitations. Key areas that remain unaddressed include the lack of cross-linguistic generalizability of ML models, limited representation of pediatric populations, and sociocultural variations in speech, which may affect the robustness and applicability of the findings. Additionally, issues such as data scarcity, inconsistent data quality, risks of model overfitting, and potential gender bias pose challenges to the development of unbiased and reliable diagnostic tools. The generalization of findings to a broader range of mental health disorders is also a concern; while Parkinson disease and schizophrenia are discussed, the exclusion of numerous other conditions limits the scope of applicability. Clarification on whether these findings can be extended to non–speech-related disorders or a recommendation for future research in this area would strengthen the manuscript.

List of Minor Concerns and Feedback

Concerns With Techniques/Analyses

- The manuscript does not thoroughly discuss model validation practices or the potential risk of bias, such as overfitting and limited sample diversity. Although the interpretations are generally sound, a more critical evaluation of the limitations of the individual studies could be included. The authors may wish to include a subsection that summarizes the validation methods used by the reviewed studies.
- There is a lack of standardization in the techniques used across the 91 studies, as most studies employ different speech tasks, which may impact the biomarkers activated or identified. Additionally, speech impairment changes with disease progression, so it would be useful to include age and more information about the disease state.
- The References section shows inconsistencies in formatting and needs to be revised to follow a uniform citation style in accordance with a journal's guidelines.
- The number of included articles is stated as 91, but Tables 1-3 present only 77 studies, while Table 4 shows 64. This discrepancy is unclear and may confuse readers. Kindly provide an explanation for the differences in the number

of articles across the tables. You can include a brief footnote in the manuscript on why those articles were excluded.
- In section 2.6 “Articles Found,” it is unclear why articles including magnetic resonance imaging, computed tomography, electroencephalogram, image, wearable sensors, video, transcription, or multimodal data were excluded. Clarify the specific scope and focus of the review that justified the exclusion of these factors.
- The year of publication listed in the table looks disorganized. The authors could reorder the studies in the table in either ascending or descending order of year of publication to help readers identify the progression of research over time.
- Please clarify why GPT-4 or GPT-4.5 (instead of GPT-3.5) was not used despite being available at the time of the study.
- Under “3. Results,” the authors could use more clarifying language while describing languages used (English was the most common language, but the results also included studies on Chinese, Greek, Spanish, Malay, and Hebrew). Since non-English language studies were excluded. It looks like they may have used studies where test sets were in different languages. Suggestion: The sentence under “3. Results” can be restructured to clarify the same.

Details for the Reproducibility of the Study

- The reproducibility of the Grading of Recommendations Assessment, Development and Evaluation (GRADE) scoring is limited due to the absence of a clearly defined rubric or framework. Provide a detailed explanation or scoring rubric highlighting how each criterion of the GRADE scoring system was applied.
- An insufficient search strategy will make it difficult for other researchers to replicate or validate the review process. Authors should expand the Method section and describe the databases used, the search terms, the inclusion or exclusion criteria, and any screening processes like PRISMA flow. This will improve the credibility and reproducibility of this study.

Figures and Tables

- Some captions lack the specific details of the dataset used, the method languages, and the clinical settings. Also, some tables are overly dense. Revise these captions to include contexts like data sources, methodology, and clinical backgrounds. The authors may consider breaking dense tables into subcategories to enhance clarity.
- The reference numbers are missing in the first column of all tables and should be added in brackets following the author names (eg, “Alan et al [23]”) to allow quick cross-referencing with the reference list.
- Not all the tables were cited within the main text of the article.
- The description of Figure 1 should be expanded further. Moreover, the authors should put the name of the primary author before the reference and the year of publication (eg, “(NAME et al (2XXX) [114]”). Figure 1 should also be revised to increase its readability. Perhaps, the authors could minimize the quadrants and increase the size of the text font.
- Divide the Participants column in Tables 1-3 into “Target Patients” and “Control Patients” to improve readability.
- It would be helpful if the tables listed the time duration of the studies.
- There are multiple spelling mistakes and excessive use of undefined abbreviations, especially in tables. There is also a lack of standardization in reporting speech features and methods, making comparison difficult.
- Could combine similar findings in each table (ie, combine cells), but keep authors’ citations in the tables.

Additional Comments

- The manuscript would benefit from figures, diagrams, or charts that summarize key trends such as ML model performance across various disorders, as well as a visual overview of the review process.
- There is insufficient detail on why speech disorders were chosen as the focal point in a rapidly expanding domain of ML-based diagnostics. Authors should add content and references to emphasize the broader relevance of ML in diagnostics and explain the reason behind their narrowing the scope to speech-based disorders.
- Number the references in order, starting with “#41.”
- In both the Abstract and Results sections, please write the abbreviation “OCD” as “obsessive-compulsive disorder (OCD).”
- In the Rationale and Results section, please revise the sentence “ML provides enables” by removing one of the verbs to correct the grammar.
- Please add a reference to the GRADE rating.
- In the Dysarthria, general section: please identify the abbreviation “PWSI-AI-AC” as “patch-wise wave splitting and integrating AI system for audio classification.”
- In the “Alzheimer’s Disease (AD)” section, please identify the abbreviation “eGeMAPS” as the “extended Geneva Minimalistic Acoustic Parameter Set.”
- “Gomez et al” should be corrected to “G´omez-Rodellar et al” in the “Parkinson’s Disease (PD)” section and Table 1.
- In the “Incorporating ML Based Speech Assessment in Clinical Practice” section, please identify “GDPR” as “General Data Protection Regulation.”
- In the Methods section, the phrase “focused on Parkinson, [3] focused on psychiatric disorders, and [4] focused on depression and suicide risk” should be revised to “focused on Parkinson, [3] on psychiatric disorders [4], and on depression and suicide risk.”
- The title includes “state of the art,” which may be misleading as the GPT-3.5-turbo model was used in this paper, and since February 27, 2025, the most current version, GPT-4.5 model, has been released. Authors should specify the model type in the title.
- Acronyms such as “CNN” and “AUC” are used without definition on page 6.
- “3.2.6 Reinke’s edemba”: It should be “edema” not “edemba.”
- This manuscript requires comprehensive proofreading and editing.

We thank the authors of the preprint for posting their work openly for feedback. We also thank all participants of the live

review call for their time and for engaging in the lively discussion that generated this review.

Acknowledgments

PREreview and JMIR Publications thank the authors of the preprint for posting their work openly for feedback. We also thank all participants of the live review for their time and for engaging in the lively discussion that generated this review.

Conflicts of Interest

VF was a facilitator of this call and one of the organizers. No other competing interests were declared by the reviewers.

References

1. Moell B, Aronsson FS, Östberg P, Beskow J. The order in speech disorder: a scoping review of state of the art machine learning methods for clinical speech classification. arXiv. Preprint posted online on Mar 3, 2025. [doi: 10.48550/arXiv.2503.04802]
2. Tricco AC, Lillie E, Zarin W, et al. PRISMA Extension for Scoping Reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med 2018 Oct 2;169(7):467-473. [doi: 10.7326/M18-0850] [Medline: 30178033]
3. Low DM, Bentley KH, Ghosh SS. Automated assessment of psychiatric disorders using speech: a systematic review. Laryngoscope Investig Otolaryngol 2020 Feb;5(1):96-116. [doi: 10.1002/lio2.354] [Medline: 32128436]
4. Cummins N, Scherer S, Krajewski J, Schnieder S, Epps J, Quatieri TF. A review of depression and suicide risk assessment using speech analysis. Speech Commun 2015 Jul;71:10-49. [doi: 10.1016/j.specom.2015.03.004]

Abbreviations

AD: Alzheimer disease
AI: artificial intelligence
GDPR: General Data Protection Regulation
GRADE: Grading of Recommendations Assessment, Development and Evaluation
ML: machine learning
PRIMSA-ScR: Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews

Edited by A Schwartz; submitted 01.05.25; this is a non–peer-reviewed article;accepted 01.05.25; published 12.05.25.

Please cite as:
Fairhurst V, Sakilay S, Mahmoud RSG, Rasania S, Moonga J, Olatoye TI, Prasad R, Venkatesan P, Medeiros V, Chalwadi UK
Peer Review of "The Order in Speech Disorder: A Scoping Review of State of the Art Machine Learning Methods for Clinical Speech Classification (Preprint)"
JMIRx Med 2025;6:e76836
URL: https://xmed.jmir.org/2025/1/e76836
doi:10.2196/76836

Peer Review of "Interactive Evaluation of an Adaptive-Questioning Symptom Checker Using Standardized Clinical Vignettes (Preprint)"

Rameshwari Prasad[1]; Prasakthi Venkatesan[2]; Shawn Asadian[3]; Randa Salah Gomaa Mahmoud[4]; Uday Kumar Chalwadi[5]; Chidi Asuzu; Benjamin Senst; J Moonga[6]; Toba Isaac Olatoye[7]

[1]University of Memphis, Memphis, TN, United States
[2]University of Illinois Urbana-Champaign, Urbana, IL, United States
[3]University of British Columbia, Vancouver, BC, Canada
[4]Zagazig University, Zagazig, Egypt
[5]Louisiana State University Health Sciences Center, Shreveport, LA, United States
[6]University College London, London, United Kingdom
[7]Kwara State Teaching Service Commission, Ilorin, Nigeria

Related Article:

Companion article: https://preprints.jmir.org/preprint/83429

(JMIRx Med 2025;6:e85624) doi:10.2196/85624

KEYWORDS

artificial intelligence; clinical decision support systems; triage; history taking; patient navigation; telemedicine; mobile apps; natural language processing; patient simulation

This is the peer-review report for the preprint "Interactive Evaluation of an Adaptive-Questioning Symptom Checker Using Standardized Clinical Vignettes."

This review is the result of a virtual, collaborative live review discussion organized and hosted by PREreview and JMIR Publications on September 18, 2025. The discussion was joined by 18 people: 2 facilitators from the PREreview team, 1 member of the JMIR Publications team, 1 author, and 14 live review participants. The authors of this review have dedicated additional asynchronous time over the course of 2 weeks to help compose this final report using the notes from the live review. We thank all participants who contributed to the discussion and made it possible for us to provide feedback on this preprint.

Summary

Artificial intelligence (AI) is rapidly transforming health care. AI has been integrated into many clinical applications, including symptom checkers that help guide users to make informed care decisions. This study [1] aimed to evaluate the triage performance and history-taking quality of an adaptive-questioning symptom checker called CareRoute. CareRoute is designed to help improve health outcomes and reduce health care costs. There were three objectives: (1) to evaluate CareRoute's triage accuracy and safety using an interactive protocol that begins with only the presenting complaint, (2) to evaluate CareRoute's ability to elicit key clinical features through adaptive questioning, and (3) to establish a reproducible methodology for evaluating the quality of history-taking symptom checkers.

With the use of 45 standardized clinical vignettes (Semigran set, *BMJ* 2015 [2]), the authors compared the platform's triage recommendations against reference standards and introduced reproducible metrics to assess history-taking quality. A physician evaluator answered CareRoute's follow-up questions. To measure the quality of history-taking, the authors introduced two new metrics: elicitation coverage and elicitation fraction. They also recorded the duration of each session and the number of questions asked. The results showed that CareRoute matched expert triage decisions in 88.9% of cases, correctly identified all emergencies with no under-triage, and used urgency-aware questioning to remain efficient. Emergency cases required fewer questions and less time, while doctor visits and self-care cases involved longer interactions.

In summary, their findings showed that CareRoute performed strongly and highlighted the importance of measuring history-taking quality when evaluating symptom checkers. This study is timely given the rapid rise of digital health tools and makes a valuable contribution by proposing a reproducible framework for evaluating adaptive-questioning tools, offering valuable insights for improving and benchmarking future digital health applications. However, reliance on a single evaluator and a modest vignette sample size limit generalizability and may not fully reflect broader real-world use. Further work is needed to validate results across users and health care contexts.

List of Major Concerns and Feedback

1. All the evaluation questions were answered by the same physician (marked as PM in the preprint), who is also one of the cofounders of the app CareRoute, meaning they are highly

familiar with its functionality. This may introduce a positive bias. One of the major questions we are left with is whether the results might have differed if additional or independent physicians had been involved in the evaluation. The authors could include additional independent evaluators or add anonymized assessments to get unbiased results.

2. The statistical tools, thresholds, and confidence intervals were not reported, which makes it difficult for others to assess or reproduce the analysis. More statistical transparency is recommended.

3. Please compare the proposed metrics of elicitation coverage and elicitation fraction with metrics proposed by other authors, such as recall rate and efficiency rate (Ben-Shabat N, Sharvit G, Meimis B, *et al*. Assessing data gathering of chatbot-based symptom checkers - a clinical vignettes study. *Int J Med Inform*. 2022 Dec:168:104897. [doi: 10.1016/j.ijmedinf.2022.104897]).

4. Please consider discussing ethical issues such as:

- What influence can the automation of triaging have on real-world health care systems? Could it replace humans? Could it misguide patients?
- Will health care systems need to adapt to triaging and history questioning apps? In what way do health care systems need to adopt to implement triaging/history questioning applications successfully?
- Could CareRoute increase or reduce the digital divide? Accessibility and inclusion—there is no mention of how accessible the tool is to people with low health literacy, disabilities, or language barriers.
- How practical is it for a person experiencing an emergency condition to interact with the CareRoute app?
- AI transparency—no mention is made of how CareRoute arrives at its triage conclusions.

5. Page 3, first paragraph: It was mentioned that "CareRoute provides four triage levels (Emergency Care, Urgent Care, Doctor Visit, Self Care), but our analysis uses a conservative 3-tier mapping that collapses Urgent Care to Doctor Visit." Why this modification was performed is not clear. Please add more clarification, as it could be the reason for the difference from the original results of [2].

6. You may consider elaborating on the strategic approach used to strengthen the internal validity of the vignette data (eg, its relevance, reliability, effectiveness, and completeness). Clarifying this would help emphasize the role of the review process in shaping and supporting the quality of the data collected and analyzed. For instance, it could be helpful to describe any systematic methods applied to prevent data saturation, as well as any techniques used to identify or remove potentially biased elements from the vignettes. Similarly, outlining the strategies used to enhance generalizability would further strengthen the study's methodological transparency. Incorporating these reflections would contribute to the overall rigor and robustness of the findings. You might find the following reference useful in framing this discussion: Spalding NJ, Phillips T. Exploring the use of vignettes: from validity to trustworthiness. *Qual Health Res*. 2007 Sep;17(7):954-62. [doi: 10.1177/1049732307306187].

List of Minor Concerns and Feedback

1. The results can be hard to follow because of the limited number of visuals and tables. It is recommended to add more visual summaries, as it would make the findings much more clear and engaging.

- Section 2.2.1 "Normalized Features: Example Mapping" could be visualized as a figure.
- Consider changing "3.4 Case Example: Kidney Stones" into a figure.

2. Some sentences in the Methods and Discussion are too long and a bit wordy. It is recommended to shorten them and add smoother transitions to make the manuscript more readable.

3. Reference 11 ("Evaluating the use of digital symptom checkers in primary care: a mixed-methods study") could not be found on the internet (Google Scholar). Please check this reference.

- There are similar articles: El-Osta A, Webber I, Alaa A, *et al*. What is the suitability of clinical vignettes in benchmarking the performance of online symptom checkers? An audit study. *BMJ Open*. 2022 Apr 27:12(4):e053566 [doi: 10.1136/bmjopen-2021-053566].
- If generative AI was used in the process of writing or for any other component of the manuscript, please declare its use.

4. Can the results be transferred to other countries or health care systems? Cultural/language bias should be considered or mentioned as a limitation—especially important to consider for global implementation.

Concluding Remarks

We thank the authors of the preprint for posting their work openly for feedback. We also thank all participants of the live review call for their time and for engaging in the lively discussion that generated this review.

Disclaimer

The authors declare that they did not use generative artificial intelligence to come up with new ideas for their review.

Conflicts of Interest

None declared.

References

1. Madda P, Kondru J. Interactive evaluation of an adaptive-questioning symptom checker using standardized clinical vignettes. JMIR Preprints. Preprint posted online on Sep 2, 2025. [doi: 10.2196/preprints.83429]
2. Semigran HL, Linder JA, Gidengil C, Mehrotra A. Evaluation of symptom checkers for self diagnosis and triage: audit study. BMJ 2015 Jul 8;351:h3480. [doi: 10.1136/bmj.h3480] [Medline: 26157077]

Abbreviations

AI: artificial intelligence

Edited by A Schwartz; submitted 10.10.25; this is a non–peer-reviewed article;accepted 10.10.25; published 24.10.25.

Please cite as:
Prasad R, Venkatesan P, Asadian S, Mahmoud RSG, Chalwadi UK, Asuzu C, Senst B, Moonga J, Olatoye TI
Peer Review of "Interactive Evaluation of an Adaptive-Questioning Symptom Checker Using Standardized Clinical Vignettes (Preprint)"
JMIRx Med 2025;6:e85624
URL: https://xmed.jmir.org/2025/1/e85624
doi:10.2196/85624

Publisher:
JMIR Publications
130 Queens Quay East.
Toronto, ON, M5A 3Y5
Phone: (+1) 416-583-2040
Email: support@jmir.org