JMIRx Med

Overlay journal for preprints with post-review manuscript marketplace
Volume 7 (2026) ISSN 2563-6316 Editor in Chief: Edward Meinert, MA (Oxon), MSc, MBA, MPA, PhD, CEng, FBCS, EUR ING

Contents

Viewpoint

Peer-Review Reports

Authors’ Response To Peer Reviewss

Original Papers

Interpreting the Estimand Framework From a Causal Inference Perspective

Jinghong Zeng[1,2], MSc

[1]Department of Statistics, University of Auckland, 38 Princes Street, Auckland, New Zealand

[2]Department of Statistics and Programming, Jiangsu Hengrui Medicine (China), Guangzhou, Guangdong, China

Corresponding Author:
Jinghong Zeng, MSc
Department of Statistics, University of Auckland, 38 Princes Street, Auckland, New Zealand

Related Articles:

Companion article: http://preprints.jmir.org/preprint/88813

Companion article: https://med.jmirx.org/2026/1/e98122

Companion article: https://med.jmirx.org/2026/1/e98125

Companion article: https://med.jmirx.org/2026/1/e98126

Companion article: https://med.jmirx.org/2026/1/e98121

Abstract

The International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use published the estimand framework in 2019. The estimand framework aims to clearly define a treatment effect for a clinical question through construction of estimands, and it has been widely applied in clinical trials in the pharmaceutical industry. The estimand framework proposes 5 attributes for an estimand: treatments, variables, target populations, population-level summaries, and intercurrent events. It also proposes the treatment policy strategy, the hypothetical strategy, the composite variable strategy, the while on treatment strategy, and the principal stratum strategy to handle intercurrent events. When people give clear definitions for these 5 attributes, they clearly define an estimand that represents a treatment effect. From a statistical perspective, a genuine or causal treatment effect is defined through a causal inference framework. This article aims to interpret the estimand framework using a causal inference framework and help researchers understand the differences between estimands and causal treatment effects. From a causal inference framework based on potential outcomes, an individual treatment effect (ITE) is defined by comparison of individual potential outcomes with experimental or control treatments, and the average treatment effect (ATE) of the experimental treatment versus the control treatment is defined as an average of all ITEs. The statistical presentation of the ATE is not equivalent to an estimand. It has the same treatments, variables, target populations, and population-level summaries as an estimand, but intercurrent events are not part of it. Intercurrent events modify the statistical presentation of the ATE through treatments, variables, and target populations, whose impact can be controlled by intercurrent event strategies. I propose that the estimand attributes can be mapped onto the statistical presentation of the ATE, and that intercurrent events act as mediation mechanisms in the attribute mapping process, which provides a novel way to incorporate the causal inference framework into the estimand framework. If the estimand framework is combined with a causal inference framework, it will gain a stronger theoretical foundation. The interpretation of the estimand framework from a causal inference perspective is useful for both industrial and academic clinical trials. Observational studies may also find useful information on causal inference theories in this article.

(JMIRx Med 2026;7:e88813) doi:10.2196/88813

KEYWORDS

causal inference; clinical trial; estimand; intercurrent event; treatment effect

Introduction

The International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use (ICH) is an international association that brings regulatory authorities and the pharmaceutical industry together to discuss the scientific and technical aspects of pharmaceutical development and develop international standardized technical guidelines that are implemented by both regulators and pharmaceutical companies [1]. It was originally founded in 1990 and restructured as a nonprofit legal entity in 2015, with regulatory and industry members such as the US Food and Drug Administration [2]. Over the past 40 years, the ICH has proposed more than 70 technical guidelines that cover different aspects of pharmaceutical development, from good clinical practice to data management and statistical analysis [2]. These guidelines provide clear, standardized requirements on datasets, statistical analysis results, and other materials, which facilitates communication between regulators and pharmaceutical companies. Companies should follow ICH guidelines when submitting drug applications to regulators. The most important guideline for statistical analysis is the Statistical Principles For Clinical Trials E9 [3]. E9 was finalized in 1998 and has become an established international statistical principle for clinical trials. In 2017, the ICH drafted the guideline E9 (R1) as an addendum to E9, and published E9 (R1) in 2019 [4]. A core advancement in E9 (R1) is the estimand framework. This framework consists of definitions of estimands and relevant terms, as well as analysis strategies. It aims to improve precision in defining a treatment effect for a clinical question. Since its release, the estimand framework has gained increasing attention in the pharmaceutical industry and academia [5-9]. Many clinical trials have used the estimand framework to develop new drugs for both oncological and nononcological diseases, and professional working groups such as the Oncology Estimand Working Group have been initiated to study how the estimand framework should be better incorporated into pharmaceutical research [10].

In the world of statistics, causal inference is used to define and estimate a genuine treatment effect [11-20]. One widely used causal inference framework is the potential outcome framework. Potential outcomes are what would happen if a participant took a control treatment or instead took another experimental treatment. Potential outcomes are not yet observed. The treatment effect on this participant is defined as some form of difference between the outcomes of the two hypothetical treatment conditions, which compares two treatments or treatment regimens. Then, the individual treatment effect can be extended to the entire population of interest. Both the estimand framework and the the causal inference framework aim to define a treatment effect. I find that the estimand framework itself does not define a treatment effect as well as it intends to. There are many important issues related to causal inference, including unmeasured confounding, noncompliance, and mediation. The estimand framework does not address these problems. It is more like a framework that makes the analysis objectives clearer. What are the relationships and differences between the estimand framework and the causal inference framework? This is my main question in this article. Drury et al [8] also studied this question, but my reasoning is quite different. Drury et al [8] introduced the potential outcome framework and a statistical formula for the estimand and then focused on defining estimands from the estimand framework and the causal inference framework in specific clinical trial examples, with the aim of discussing how the two frameworks are linked. They argued that the two frameworks do not compete. I strongly agree with this. However, they did not explain how the statistical formula for the estimand is developed in the potential outcome framework and how this formula is related to practical statistical analysis. They also did not discuss in detail how intercurrent events and different strategies would affect the causal interpretation, as intercurrent events are a major part of the estimand framework. From my point of view, Drury et al [8] have not yet grasped the nature of the connections and differences. In this article, I will describe in detail how a causal inference framework can be developed based on potential outcomes and how the statistical presentation of a genuine treatment effect can be developed and related to statistical models people actually work with. The causal inference framework comes from my recent paper in *Statistics in Medicine* [11]. It can deal with several important problems in causal inference, such as unmeasured confounding and noncompliance. It is a practical, useful solution for more complex situations. I will also compare the attributes of the estimand with the statistical formula for a treatment effect, and I will discuss in detail how different intercurrent event strategies affect the causal interpretation of a treatment effect. Finally, I will propose a novel way to incorporate the causal inference framework into the estimand framework through attribute mapping.

I would like to explain a little why causal interpretation is needed by discussing two concepts, the intention-to-treat (ITT) principle and clinical significance. First, the estimand framework heavily relies on the ITT principle, or at least, clinical implementations heavily rely on the ITT principle [21-29]. The E9 states that the ITT principle “asserts that the effect of a treatment policy can be best assessed by evaluating on the basis of the intention to treat a subject (ie, the planned treatment regimen) rather than the actual treatment given. It has the consequence that subjects allocated to a treatment group should be followed up, assessed and analysed as members of that group irrespective of their compliance to the planned course of treatment” [3]. The ITT principle is the gold standard in clinical trials, and it usually relies on the benefits of randomization. Randomization creates balance between treatment groups and helps reduce bias from unmeasured confounding and noncompliance in estimation of treatment effects. Here, I discuss unmeasured confounding. Confounders are factors that affect both treatments and endpoints. For example, in a clinical trial that studies whether a new drug can reduce blood pressure, older participants may absorb the drug with more difficulty and already have higher blood pressure. Hence, age affects both the new drug and blood pressure, and thus age is a confounder. If confounders are known and measured in a study, then they are “measured confounders.” If they are unknown to the researchers, or known but not measured in a study, then they are “unmeasured confounders.” In the previous example, if age is not collected in the trial, then age is an unmeasured confounder. Unmeasured confounding is an important source of bias in

estimation of a causal treatment effect [11,30-33]. If it is not adjusted for in statistical analysis, treatment effect estimation can be biased. In the previous example, if age is not adjusted for in statistical analysis, the new drug effect is likely to be underestimated because it is likely to lead to smaller decreases in blood pressure among older people. Randomization, due to its nature, is not confounded with treatments and endpoints. Hence, even if unmeasured confounders exist between treatments and endpoints, the ITT principle uses randomization to adjust for it, where estimation of the treatment effect will not be affected by unmeasured confounding. However, under the ITT principle, what is estimated in statistical analysis is not the genuine effect of the treatment of interest, but the effect of the random assignment. If a significant effect can be detected under the ITT principle, it means that the treatment effect of interest is also significant, because randomization affects endpoints only through treatments of interest; however, the genuine magnitude of the treatment effect will not be known.

Second, I distinguish two concepts: statistical significance and clinical significance [34-42]. Statistical significance typically indicates that the differences between groups being compared are significant in statistical hypothesis tests. Clinical significance typically indicates that the differences between groups being compared are significant from a physician's perspective. The two kinds of significance are different. People rely on statistical significance to provide evidence that a new drug outperforms a placebo or active comparator, while physicians use medical knowledge to judge the actual clinical benefits of a new drug. For example, a new drug might reduce systolic blood pressure by 2 mm Hg compared to a placebo, and this difference might be statistically significant in hypothesis testing, but a physician might still advise that a difference under 5 mm Hg would not sufficiently improve a patient's conditions, judging that a difference of 2 mm Hg is not clinically significant. In order to better judge clinical significance, it is necessary to more accurately quantify the treatment effect. This idea also works in other areas of pharmaceutical development, such as precision medicine. Knowing the genuine treatment effect helps compare subgroup differences more efficiently. This does not mean that current analysis approaches, such as the ITT principle, are bad. Many effective drugs have been developed with these approaches. However, as technology and medical demands evolve, analysis approaches may also have to evolve to match.

Let me introduce some estimand-related concepts before making comparisons. E9 (R1) includes a glossary , which I've provided in Multimedia Appendix 1 [4]. The estimand framework proposes estimands and distinguishes them from estimators and estimates with regard to statistical roles in the estimation of treatment effects. An estimand is a precise definition of a treatment effect in a clinical question. An estimator is a statistical method that estimates the estimand, and an estimate is a result from the estimator. The estimand framework introduces 5 attributes for an estimand. They are treatments, variables, target populations, population-level summaries, and intercurrent events [4]. The 5 attributes together define an estimand. Generally, treatments are drugs used in clinical trials. They can be new drugs or new combinations of drugs. Variables are outcomes used to assess efficacy and safety of treatments, such as blood pressure and the occurrence rate of adverse events. They are also called endpoints. A target population is a group of people that satisfy specific conditions of clinical interest, such as people older than 60 years with hypertension. A population-level summary is a statistical approach to compare the treatment effect among different groups, such as the risk difference between the treatment arm and the placebo arm. Intercurrent events are events "occurring after treatment initiation that affect either the interpretation or the existence of the measurements associated with the clinical question of interest" [4]. Usually, they affect the definitions of treatments, variables, and target populations. Intercurrent events are very common in practice, including use of concomitant therapies, treatment switch, and death before endpoint measurement [4]. For example, when concomitant therapies are used, their effects are mixed with the treatment effect of interest, which may bias estimation of the treatment effect. When a participant in the treatment arm switches to the placebo arm, the drug effect on this participant no longer comes from the original treatment. When a participant dies before an endpoint assessment, the endpoint will become missing.

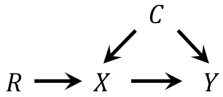
The estimand framework proposes 5 strategies to handle intercurrent events, with consideration of study objectives [4]. They are the treatment policy strategy, the hypothetical strategy, the composite variable strategy, the while on treatment strategy, and the principal stratum strategy [4]. The treatment policy strategy considers intercurrent events as part of the treatments being compared in clinical trials. The hypothetical strategy hypothesizes what would happen if no intercurrent event occurred. The composite variable strategy considers intercurrent events as part of the variables being assessed in clinical trials. The while on treatment strategy excludes any data after intercurrent events, including treatments and endpoints after intercurrent events. The principal stratum strategy considers the treatment effect in specific subpopulations of the entire target population. To illustrate the use of estimands in the real world, I have described estimands from some recent clinical trials in Multimedia Appendix 2 [43-48].

A Causal Inference Framework Compared to Estimands

I would like to introduce a causal inference framework based on potential outcomes I have previously reported [11]. The framework described here is simplified: noncompliance and many important assumptions are not mentioned; only two treatment conditions are considered. However, a simplified framework is good for a broad audience when the definition of a treatment effect is accurate.

Suppose there is a 2-arm randomized controlled clinical trial, with full compliance to treatment. This clinical trial has a sample size of N, a binary treatment X, a continuous endpoint Y, a randomization scheme R, and confounders C. R only affects X. X only affects Y. Hence, R affects Y only through X. C affects both X and Y. Their causal relationships can be shown in a causal directed acyclic graph (Figure 1).

Figure 1. Causal directed acyclic graph for the treatment *X*, the endpoint *Y*, the randomization scheme *R*, and the confounders *C*.



X, Y, R are random vectors of length N. C includes both measured and unmeasured confounders. It is a random matrix of row dimension N. Xi, Yi, Ri, Ci represent random variables or vectors for the participant i, where i∈ 1,2,3,…,N. Ri=0 means that the participant is assigned to the control arm, and Ri=1 means that the participant is assigned to the treatment arm. XiRi=0=0 means that the participant takes the control treatment if assigned to the control arm, and XiRi=1=1 means that the participant takes the experimental treatment that is of primary clinical interest if assigned to the treatment arm. YiXiRi=0=0 is the endpoint if the participant takes the control treatment as assigned, and YiXiRi=1=1 is the endpoint if the participant takes the experimental treatment as assigned. Ri, Xi and Yi are potential outcomes.

The clinical goal is to estimate the overall treatment effect on the endpoint of taking the experimental treatment versus taking the control treatment among all participants. For the participant i, if they have XiRi=0=0, then they have YiXiRi=0=0, and if they have XiRi=1=1 , then they have YiXiRi=1=1. For this participant, the individual treatment effect (ITE) is defined as the difference between two potential outcomes of Yi. That is,

$$ITE=Yi(Xi(Ri=1)=1)-Yi(Xi(Ri=0)=0).$$

I further assume that the distribution of Y has linear functional forms with X and C. I also assume that the ITE is same for all participants. These assumptions can be relaxed theoretically [11]. The distribution of Y is thus assumed to be

$$Y=\beta0+\beta1X+\beta2C+\varepsilon.$$

Then, for the participant i,

$$Yi(Xi(Ri=1)=1)=\beta0+\beta1+\beta2Ci+\varepsilon i,$$

$$YiXiRi=0=0=\beta0+\beta2Ci+\varepsilon i.$$

Hence, β1=YiXiRi=1=1-YiXiRi=0=0, and it equals the ITE for the participant i. The ITE indicates how the endpoint would change when only the treatment condition changes for this participant. There are N ITEs. Suppose a researcher is interested in an average treatment effect (ATE) of taking the experimental treatment versus taking the control treatment among all participants, which is the answer to the clinical goal mentioned above. The ATE is an average of all ITEs. It is defined as

$$ATE=EYXR=1=1-YXR=0=0.$$

This indicates that β1 also equals the ATE, because ATE=Eβ1=β1. Under this specific distributional assumption of Y, β1 acts as an estimator to the ATE. If the researcher makes different distributional assumptions, they will obtain different estimators.

Further, based on expectation properties, it is clear that

$$ATE=E(Y(X(R=1)=1))-E(Y(X(R=0)=0)).$$

This also means that the estimand is the difference between the average outcome of taking the experimental treatment among all participants and the average outcome of taking the control treatment among all participants. The problem is that, in the real world, each participant only takes one kind of treatment, and only one of the two hypothetical situations with regard to two arms can happen. In this clinical trial setting, for each participant, the observed random assignment Rio is either 0 or 1, the observed treatment Xio is either XiRi=0 or XiRi=1, and the observed outcome Yio is either YiXiRi=0 or YiXiRi=1. The relationships between potential outcomes and observed variables can be described as

$$Xio=Xi(Ri=Rio),$$

$$Yio=Yi(Xi(Ri=Rio)),$$

which implies that Yo=Y(Xo). Through Yo=Y(Xo) and the distribution of Y, the researcher has a causal linear model of observed variables to estimate the estimand. The linear model is given as

$$Yo=\beta0+\beta1Xo+\beta2C+\varepsilon,$$

$$E\varepsilon=0,Var\varepsilon=\sigma2.$$

The linear model is the statistical model that the researcher builds using actual clinical trial data, where β1 is not changed. After the linear model is built with data, an estimate β^1 on β1 would be obtained. β^1 is an estimate of the ATE.

The statistical formula of the ATE as E(YXR=1=1-YXR=0=0) already contains 4 attributes for an estimand, but it omits the attribute of intercurrent events. The attribute of treatments is represented by X. The attribute of variables is represented by Y. The attribute of population-level summaries is represented by the difference in the ATE between taking the experimental treatment and taking the control treatment among all participants, that is, the expectation form. The attribute of target population is implicitly stated as the clinical trial population and can be made explicit in the definition of the estimand. If a selection variable S indicates how the target population is selected from the general public, the formula can be updated to EYXR=1=1-YXR=0=0S.

On the other hand, the 4 attributes of treatments, variables, population-level summaries, and target populations are not sufficient to define a clear treatment effect without the statistical formula for the ATE. The estimand framework raises awareness of a genuine treatment effect and makes progress on estimand attributes related to a treatment effect. If it can be improved with a clear statistical definition, it will gain a stronger theoretical foundation.

Let us take a look at the ITT principle. Under the ITT principle, a linear model of Yo that does not consider auxiliary variables such as measured variables in C could be

$$Yo=\beta0+\beta1Ro+\varepsilon,$$

$$E\varepsilon=0,Var\varepsilon=\sigma2,$$

where β1 now is an estimator of the average effect of the random assignment, or in other words, the average randomization effect (ARE). Note that not including auxiliary variables does not make this model invalid, but including auxiliary variables may

improve model efficiency. The statistical formula for the ARE is EYR=1-YR=0, from which the ARE is the difference between the average outcome of being randomized to the treatment arm among all participants and the average outcome of being randomized to the control arm among all participants. Hence, the ARE is different from the ATE. The ARE does not represent a genuine effect of the experimental treatment compared to the control treatment. It is the randomization effect, or specifically, the effect of being randomized to an experimental arm versus being randomized to a control arm. When the estimand framework is used with the ITT principle, usually the researcher will obtain an estimate for the ARE rather than the ATE. It also indicates that the estimand framework may depend on the statistical analysis approach to define a "treatment" effect.

Impact of Intercurrent Event Strategies on Treatment Effects

I would like to summarize how intercurrent events and their strategies are related to the ATE. I provide details for each intercurrent event strategy in Multimedia Appendix 3.

The treatment policy strategy includes intercurrent events in the definition of treatments. It defines a new treatment and a new endpoint. The new treatment is a combination of the original treatment and other treatments, such as the original treatment plus rescue therapy. It is also called a treatment policy [4]. The endpoint is affected by the new treatment, so it is different from the original endpoint. The ATE is changed to a genuine treatment effect at the endpoint of the experimental treatment policy versus the control treatment policy.

The hypothetical strategy hypothesizes nonexistence of intercurrent events and makes relevant data missing. Data missingness can be applied to the variable or the treatment, but it does not change the definition of the treatment and the variable, which means that the ATE will not be changed. Hence, this strategy can maintain the original treatment effect. It is more like a statistical imputation approach because suitable statistical methods should be used to impute missing data.

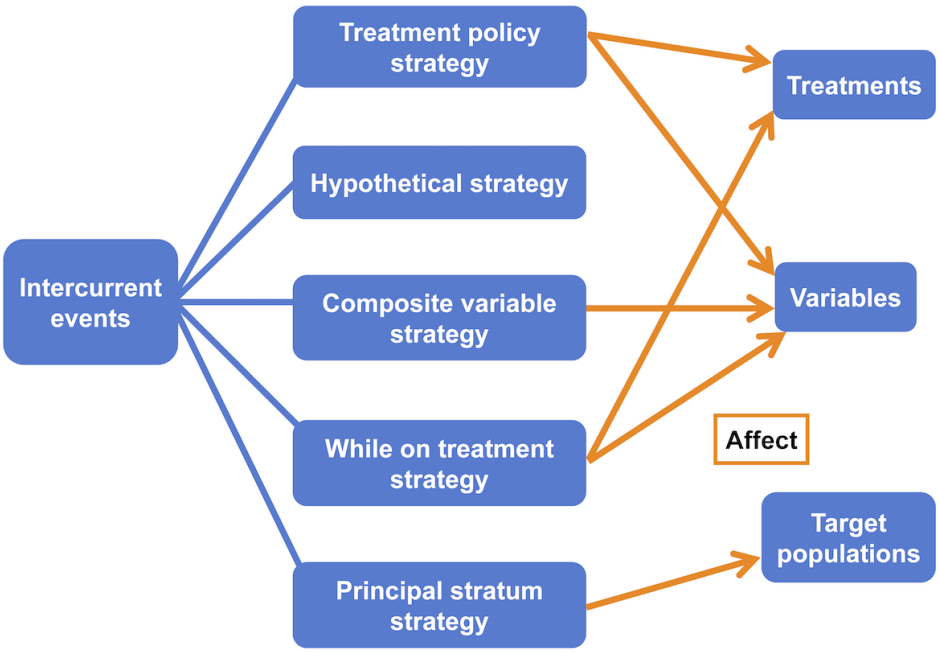
The composite variable strategy includes intercurrent events in the definition of endpoints. It defines a new composite endpoint that consists of both the original endpoint and a new component. For example, death could be a new component for an endpoint that measures serious adverse events [48]. Usually, the composite endpoint is broader than the original endpoint, and it is necessary to explain the rationale for the new component. The ATE is changed to a genuine treatment effect on the composite endpoint of the experimental treatment versus the control treatment.

The while on treatment strategy uses available data measured before intercurrent events and discards any data after the events. It defines a new treatment and a new endpoint by modifying the observation time of the original treatment and endpoint [4]. Intercurrent events usually make the observation time shorter than the planned duration, and the new treatment and endpoint last up to the occurrence of intercurrent events. The ATE is changed to a genuine treatment effect of the experimental treatment versus the control treatment before intercurrent events occur.

The principal stratum strategy estimates the ATE in a subpopulation of participants who would experience intercurrent events or a subpopulation who would not experience intercurrent events, instead of the original target population. Each subpopulation is a principal stratum [4]. For each participant, the treatment and the variable are not changed. That is, the ITE for each participant is not changed. The ATE becomes an average of all ITEs in the principal stratum. Hence, the ATE could be changed, especially under an assumption that the ITE is not identical for all participants.

Intercurrent events indirectly affect the ATE by directly affecting the treatments, the variables, the target populations, or some of these estimand attributes at the same time. Intercurrent event strategies can control the impact of intercurrent events. They can direct the impact toward certain estimand attributes, or even stop the impact, as shown in Figure 2. For example, suppose an intercurrent event is discontinuation of trial treatment and it affects both the treatment and the variable [47]. The treatment policy strategy includes treatment discontinuation in the treatment policy and changes the treatment and the variable, while the hypothetical strategy makes data after treatment discontinuation missing and stops the impact of treatment discontinuation. Once we know how these strategies work, any further strategy could be interpreted similarly with regards to the treatments, the variables, and the target populations.

Figure 2. Different ways intercurrent event strategies control the impact of intercurrent events.

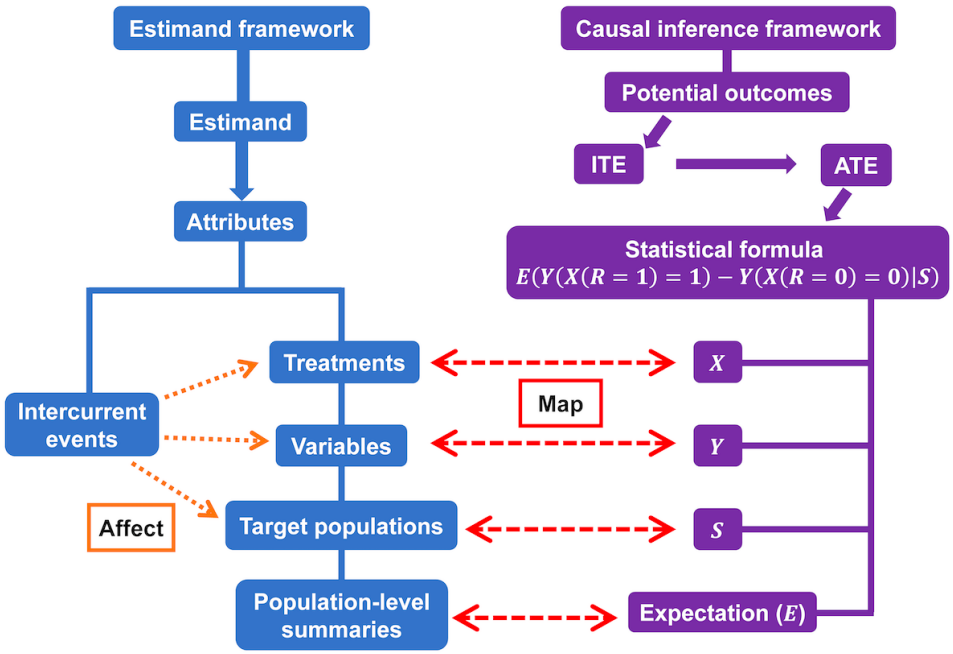


Recommendations and Future Outlook

The estimand framework serves one more pragmatic purpose: it informs industry nonstatisticians of a precise definition of a treatment effect [8]. This practical value is highly appreciated, but we need to ask, "How can precision be guaranteed?" A theoretical foundation will not only guarantee a precise definition of treatment effects, but it will also guarantee a precise understanding of treatment effects. For many years, regulatory authorities and industry professionals have made great efforts to ensure good practice related to the estimand framework. However, this does not mean that every professional knows exactly what the estimand framework is doing. If professionals could be educated with an introduction to causal inference so that they understand what a treatment effect is, how intercurrent events affect it, and the impact of various intercurrent event strategies on treatment effects, it might be possible to establish an industry consensus on a precise understanding of treatment effects. This would make communication between statisticians and nonstatisticians less time-consuming, and planning and analysis of clinical trials would improve.

Hence, it will be useful to incorporate a causal inference framework into the estimand framework. The first question is, "Which causal inference framework should professionals use?" The potential outcome framework is a good start. It provides a clear definition of a genuine treatment effect. Some people may argue that choosing a causal inference framework is more related to data analysis. From the causal inference framework, we can see that $\beta 1$ in the linear model actually comes from construction of distributions for potential outcomes. A causal inference framework is a basis for subsequent statistical models, including linear models and more complex modeling approaches. Next, how could the causal inference framework be used in the estimand framework? I provide a novel way to map estimand attributes onto the statistical presentation of the ATE, and intercurrent events act as mediation mechanisms in the attribute mapping process, as shown in Figure 3. The treatment, the endpoint, the population and the expectation form from the statistical presentation of the ATE correspond in order to the treatment, the variable, the target population and the population-level summary from the estimand attributes. Intercurrent events themselves are not part of the statistical presentation of the ATE, but they modify the statistical presentation through treatments, variables and target populations. It implies that the ATE is not equivalent to the estimand, but the two concepts share core attributes, which provides a link between the estimand framework and the causal inference framework. Similar to Drury et al [8], I argue that the estimand framework and the causal inference framework are not "competing." In addition, this link also indicates a possibility that the estimand framework may borrow theoretical benefits from the causal inference framework.

Figure 3. The attribute mapping process to connect the estimand framework and the causal inference framework. ATE: average treatment effect; ITE: individual treatment effect.



Interpretation of the estimand framework from a causal inference perspective is useful for clinical trials both in the pharmaceutical industry and in academia. When more and more academic clinical trials conduct analyses using the standardized estimand framework, communication about treatment effect results between the industry and academia may be greatly facilitated, which could improve the efficiency of clinical trials in a broad sense.

Further work is needed. For example, Figure 3 might have to be transformed into a format that professionals prefer. More complex situations, such as unmeasured confounding, noncompliance, and multiple treatment conditions, should be discussed. The causal inference framework discussed above can be expanded to deal with these issues. In addition to risk difference, more population-level summaries should also be discussed, such as relative risk and odds ratio. A review of statistical principles and models might be necessary to understand analytical performance under the estimand framework updated with causal inference theories.

Funding

There was no funding for this work.

Data Availability

No data was used in this work.

Conflicts of Interest

None declared.

Multimedia Appendix 1
A glossary of estimand-related concepts from E9 (R1).
[PDF File, 59 KB - xmed_v7i1e88813_app1.pdf]

Multimedia Appendix 2
Examples of estimands from recent clinical trials.
[PDF File, 78 KB - xmed_v7i1e88813_app2.pdf]

Multimedia Appendix 3
Impact of intercurrent event strategies on treatment effects.
[PDF File, 179 KB - xmed_v7i1e88813_app3.pdf]

References

1. International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use. URL: https://www.ich.org/ [accessed 2023-07-15]
2. The International Council for Harmonisation: an overview. International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use. 2025. URL: https://admin.ich.org/sites/default/files/2025-12/OverviewOfICH_2025_1126_0.pdf [accessed 2026-02-17]
3. E9 - Statistical principles for clinical trials. International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use. 1998. URL: https://www.ich.org/page/efficacy-guidelines#9-1 [accessed 2023-10-30]
4. E9 (R1) Addendum: statistical principles for clinical trials. International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use. 2019. URL: https://www.ich.org/page/efficacy-guidelines#9-2 [accessed 2023-10-30]
5. Pohl M, Baumann L, Behnisch R, Kirchner M, Krisam J, Sander A. Estimands—a basic element for clinical trials. Part 29 of a series on evaluation of scientific publications. Dtsch Arztebl Int 2021:883-888. [doi: 10.3238/arztebl.m2021.0373]
6. Kahan BC, Hindley J, Edwards M, Cro S, Morris TP. The estimands framework: a primer on the ICH E9(R1) addendum. BMJ 2024 Jan 23;384:e076316. [doi: 10.1136/bmj-2023-076316] [Medline: 38262663]
7. Heinrich M, Zagorscak P, Bohn J, Knaevelsrud C, Schulze L. Using the ICH estimand framework to improve the interpretation of treatment effects in internet interventions. NPJ Digit Med 2025 Aug 20;8(1):535. [doi: 10.1038/s41746-025-01936-0] [Medline: 40835721]
8. Drury T, Bartlett JW, Wright D, Keene ON. The estimand framework and causal inference: complementary not competing paradigms. Pharm Stat 2025;24(5):e70035. [doi: 10.1002/pst.70035] [Medline: 40847780]
9. Lanius V, Glocker B, Lösch C, et al. Realizing the benefits of the estimand framework when reporting and communicating clinical trial results-some recommendations. Trials 2025 Jul 11;26(1):241. [doi: 10.1186/s13063-025-08915-6] [Medline: 40640873]
10. Oncology Estimand Working Group. URL: https://oncoestimand.github.io/oncowg_webpage/docs/ [accessed 2024-05-30]
11. Zeng J. A Bayesian approach to estimate causal average treatment effects under unmeasured confounding. Stat Med 2026 Mar;45(6-7):e70461. [doi: 10.1002/sim.70461] [Medline: 41761686]
12. Rubin DB. Estimating causal effects of treatments in randomized and nonrandomized studies. J Educ Psychol 1974;66(5):688-701. [doi: 10.1037/h0037350]
13. Rubin DB. Bayesian inference for causal effects: the role of randomization. Ann Statist 1978;6(1):34-58. [doi: 10.1214/aos/1176344064]
14. Holland PW. Statistics and causal inference. J Am Stat Assoc 1986 Dec;81(396):945-960. [doi: 10.1080/01621459.1986.10478354]
15. Imbens GW, Angrist JD. Identification and estimation of local average treatment effects. Econometrica 1994 Mar;62(2):467. [doi: 10.2307/2951620]
16. Imbens GW, Rubin DB. Bayesian inference for causal effects in randomized experiments with noncompliance. Ann Statist 1997;25(1):305-327. [doi: 10.1214/aos/1034276631]
17. Rubin DB. Causal inference using potential outcomes: design, modeling, decisions. J Am Stat Assoc 2005;100(469):322-331. [doi: 10.1198/016214504000001880]
18. Vandenbroucke JP, Broadbent A, Pearce N. Causality and causal inference in epidemiology: the need for a pluralistic approach. Int J Epidemiol 2016 Dec 1;45(6):1776-1786. [doi: 10.1093/ije/dyv341] [Medline: 26800751]
19. Weed DL. Commentary: causal inference in epidemiology: potential outcomes, pluralism and peer review. Int J Epidemiol 2016 Dec 1;45(6):1838-1840. [doi: 10.1093/ije/dyw229] [Medline: 28130322]
20. VanderWeele TJ. On causes, causal inference, and potential outcomes. Int J Epidemiol 2017(6):dyw230. [doi: 10.1093/ije/dyw230]
21. McCoy CE. Understanding the intention-to-treat principle in randomized controlled trials. West J Emerg Med 2017 Oct;18(6):1075-1078. [doi: 10.5811/westjem.2017.8.35985] [Medline: 29085540]
22. Nagel S, Haussen DC, Nogueira RG. Importance of the intention-to-treat principle. JAMA Neurol 2020 Jul 1;77(7):905-906. [doi: 10.1001/jamaneurol.2020.0848] [Medline: 32364570]
23. Sicklick JK, Kato S, Okamura R, Kurzrock R. Precision oncology: the intention-to-treat analysis fallacy. Eur J Cancer 2020 Jul;133(133):25-28. [doi: 10.1016/j.ejca.2020.04.002] [Medline: 32422506]
24. Tripepi G, Chesnaye NC, Dekker FW, Zoccali C, Jager KJ. Intention to treat and per protocol analysis in clinical trials. Nephrology (Carlton) 2020 Jul;25(7):513-517. [doi: 10.1111/nep.13709] [Medline: 32147926]
25. Santos-Gallego CG, Requena-Ibanez JA, Badimon J. Per-protocol versus intention-to-treat in clinical trials: the example of GLOBAL-LEADERS trial. J Am Heart Assoc 2022 May 17;11(10):e025561. [doi: 10.1161/JAHA.122.025561] [Medline: 35574954]
26. Ahn EJ, Kang H. Intention-to-treat versus as-treated versus per-protocol approaches to analysis. Korean J Anesthesiol 2023 Dec;76(6):531-539. [doi: 10.4097/kja.23278]
27. Morga A, Latimer NR, Scott M, Hawkins N, Schlichting M, Wang J. Is intention to treat still the gold standard or should health technology assessment agencies embrace a broader estimands framework?: Insights and perspectives from the National Institute for Health and Care Excellence and Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen on

the International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use E9 (R1) Addendum. Value Health 2023 Feb;26(2):234-242. [doi: 10.1016/j.jval.2022.08.008] [Medline: 36150999]
28. Armijo-Olivo S, Barbosa-Silva J, de Castro-Carletti EM, et al. Intention-to-treat analysis in clinical research: basic concepts for clinicians. Am J Phys Med Rehabil 2024 Sep 1;103(9):845-857. [doi: 10.1097/PHM.0000000000002444] [Medline: 38320245]
29. Molero-Calafell J, Burón A, Castells X, Porta M. Intention to treat and per protocol analyses: differences and similarities. J Clin Epidemiol 2024 Sep;173:111457. [doi: 10.1016/j.jclinepi.2024.111457] [Medline: 38977160]
30. VanderWeele TJ, Arah OA. Bias formulas for sensitivity analysis of unmeasured confounding for general outcomes, treatments, and confounders. Epidemiology (Sunnyvale) 2011;22(1):42-52. [doi: 10.1097/EDE.0b013e3181f74493]
31. Dorie V, Harada M, Carnegie NB, Hill J. A flexible, interpretable framework for assessing sensitivity to unmeasured confounding. Stat Med 2016 Sep 10;35(20):3453-3470. [doi: 10.1002/sim.6973] [Medline: 27139250]
32. Groenwold RHH, Shofty I, Miočević M, van Smeden M, Klugkist I. Adjustment for unmeasured confounding through informative priors for the confounder-outcome relation. BMC Med Res Methodol 2018 Dec 22;18(1):174. [doi: 10.1186/s12874-018-0634-3] [Medline: 30577773]
33. Gaster T, Eggertsen CM, Støvring H, Ehrenstein V, Petersen I. Quantifying the impact of unmeasured confounding in observational studies with the E value. BMJ Med 2023;2(1):e000366. [doi: 10.1136/bmjmed-2022-000366] [Medline: 37159620]
34. Ogles BM, Lunnen KM, Bonesteel K. Clinical significance: history, application, and current practice. Clin Psychol Rev 2001 Apr;21(3):421-446. [doi: 10.1016/s0272-7358(99)00058-6] [Medline: 11288608]
35. Man-Son-Hing M, Laupacis A, O'Rourke K, et al. Determination of the clinical importance of study results. J Gen Intern Med 2002 Jun;17(6):469-476. [doi: 10.1046/j.1525-1497.2002.11111.x] [Medline: 12133163]
36. Sharma H. Statistical significance or clinical significance? A researcher's dilemma for appropriate interpretation of research results. Saudi J Anaesth 2021;15(4):431-434. [doi: 10.4103/sja.sja_158_21] [Medline: 34658732]
37. Kaminsky DA, Simpson SJ, Berger KI, et al. Clinical significance and applications of oscillometry. Eur Respir Rev 2022 Mar 31;31(163):210208. [doi: 10.1183/16000617.0208-2021] [Medline: 35140105]
38. Bloom DA, Kaplan DJ, Mojica E, et al. The minimal clinically important difference: a review of clinical significance. Am J Sports Med 2023 Feb;51(2):520-524. [doi: 10.1177/03635465211053869] [Medline: 34854345]
39. Boccardi V, Orr ME, Polidori MC, Ruggiero C, Mecocci P. Focus on senescence: clinical significance and practical applications. J Intern Med 2024 May;295(5):599-619. [doi: 10.1111/joim.13775] [Medline: 38446642]
40. Elasan S. The difference between clinical significance and statistical significance: an important distinction for clinical research. Turk J Med Sci 2024;54(6):1419. [doi: 10.55730/1300-0144.5925] [Medline: 39734353]
41. Qin Z, Zhu Y, Shi DD, Chen R, Li S, Wu J. The gap between statistical and clinical significance: time to pay attention to clinical relevance in patient-reported outcome measures of insomnia. BMC Med Res Methodol 2024 Aug 8;24(1):177. [doi: 10.1186/s12874-024-02297-0] [Medline: 39118002]
42. Mohanty CR, Barik AK, David GAJ, Radhakrishnan RV. Clinical significance versus statistical significance: does it matter in clinical practice? Indian J Anaesth 2025 Feb;69(2):251-252. [doi: 10.4103/ija.ija_1151_24] [Medline: 40160911]
43. Eli Lilly and Company. A study of dulaglutide (LY2189265) in participants with type 2 diabetes mellitus (AWARD-10). ClinicalTrials.gov. 2015. URL: https://clinicaltrials.gov/study/NCT02597049 [accessed 2026-02-20]
44. Novo Nordisk A/S. Effect and safety of liraglutide 3.0 mg as an adjunct to intensive behavior therapy for obesity in a non-specialist setting (SCALE IBT). ClinicalTrials.gov. 2016. URL: https://clinicaltrials.gov/study/NCT02963935 [accessed 2026-02-20]
45. Hoffmann-La Roche. A study to investigate the safety, tolerability, pharmacokinetics, pharmacodynamics and efficacy of risdiplam (RO7034067) in type 2 and 3 spinal muscular atrophy (SMA) participants (SUNFISH). ClinicalTrials.gov. 2016. URL: https://clinicaltrials.gov/study/NCT02908685 [accessed 2026-02-20]
46. GlaxoSmithKline. Comparative study of ELLIPTA dry powder inhaler (DPI) versus DISKUS DPI used with handihaler DPI in subjects with chronic obstructive pulmonary disease (COPD). ClinicalTrials.gov. 2017. URL: https://clinicaltrials.gov/study/NCT03227445 [accessed 2026-02-20]
47. ALK-Abelló A/S. A study in children and adolescents with birch pollen-induced rhinoconjunctivitis (treetop). ClinicalTrials.gov. 2021. URL: https://clinicaltrials.gov/study/NCT04878354 [accessed 2026-02-20]
48. Albert B. Sabin Vaccine Institute. Monovalent chimpanzee adenoviral-vectored Marburg virus vaccine in healthy adults. ClinicalTrials.gov. 2023. URL: https://clinicaltrials.gov/study/NCT05817422 [accessed 2026-02-20]

Abbreviations

ARE: average randomization effect
ATE: average treatment effect
ICH: International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use
ITE: individual treatment effect
ITT: intention-to-treat

Edited by A Schwartz; submitted 02.Dec.2025; peer-reviewed by H Wu, L Zhang, Q Zhang; revised version received 05.Apr.2026; accepted 08.Apr.2026; published 22.May.2026.

Please cite as:
Zeng J
Interpreting the Estimand Framework From a Causal Inference Perspective
JMIRx Med 2026;7:e88813
URL: https://xmed.jmir.org/2026/1/e88813
doi:10.2196/88813

Peer Review of “Automating Individualized Notification of Drug Recalls to Patients: Complex Challenges and Qualitative Evaluation”

Alissa Russ, PhD, PharmD

Purdue University, 575 Stadium Mall Drive, West Lafayette, IN, United States

Related Articles:

Companion article: https://www.medrxiv.org/content/10.1101/2024.09.18.24312141v1

Companion article: https://med.jmirx.org/2026/1/e82609

Companion article: https://med.jmirx.org/2026/1/e68345

(JMIRx Med 2026;7:e82613) doi:10.2196/82613

KEYWORDS

notification system; drug recalls; patient safety; medication; electronic health records; prescriptions; decision support

This is a peer review report for “Automating Individualized Notification of Drug Recalls to Patients: Complex Challenges and Qualitative Evaluation.”

Round 1 Review

General Comments

This manuscript [1] describes interesting and novel work with far-reaching patient safety implications. The authors developed an automated system in the electronic health record (EHR) of an academic medical center that scans for drug recalls, matches up National Drug Codes of recalled medication on a patient’s medical list, and sends notifications through the EHR portal to the patient, providing them with more information on the recall. The authors then conducted a qualitative analysis of 9 patients’ perceptions of a fictious recall notice. Despite successful development of the automated system, many limitations prevented the widescale adoption of this system in 2 clinics associated with the large academic medical center. The outcome of the work—a decision was made not to deploy the new software for drug recalls—was surprising, and it is important that “failed” implementation work also be published. That said, key weaknesses of the manuscript are the lack of important details, need for better organization of the content, and the need for much stronger scientific and technical writing to accurately interpret the methods, results, and implications. These weaknesses also made it much more difficult to read and evaluate the manuscript. Despite the importance of the topic, the small sample size of patients also limits the work’s impact.

Specific Comments

Title

It would be helpful if the title were a bit more specific about the technology, study methods (qualitative), and notification recipients (patients, providers, etc).

Abstract

1. The Background section appears to be contradictory. Sentence 2 says the Food and Drug Administration has ways to notify health care professionals (HCPs) and patients, but then the following sentences seem to say the opposite.
2. A few more details here on the type of platform would be helpful…software app? Web-based platform, etc? And what are the intended user types? (HCPs and patients? Or just patients?)
3. The choice of methods doesn’t seem to follow the Background section. Why was it necessary to include the clinics, rather than just work directly with the patients? Or, why was the focus on clinics, rather than pharmacies? (These comments apply to the main Introduction and Methods sections, as well.)
4. I expected the “program description” to appear in the Methods section, not the Results.

Introduction

1. The second and third sentences of the first paragraph of the Introduction: any studies or references to back up this claim?
2. No information is included on if/what literature explores this or similar topics.
3. I would recommend adding more information on the process pharmacies currently have in place for notifying patients of recalls. Also add any literature that exists showing how often patients then contact their providers or add quantitative data to highlight this extra burden on providers to emphasize the problem.
4. I expected the funding information in the last sentence of the first paragraph to be included in a funding statement or the acknowledgments (rather than the Introduction) and the rest of that statement to be described in the Methods.

Setting

1. I expected this to appear under a larger Methods section.
2. What was the goal sample size and rationale for the sample size? There is missing demographic information on the participating patients.
3. So the Fast Healthcare Interoperability Resources (FHIR) portion notified HCPs? The intended recipients are not specified for that part of the program.
4. "EHR build" was unexpected as a reader. Is that a third part? How does it fit into the first 2 parts?
5. The screenshots and figures are useful.
6. Even for a convenience sample, more details are needed on recruitment. How did you choose which patients to email? How many were emailed for recruitment? Were patients emailed and recruited sequentially, for example? Were there any exclusion or inclusion criteria for patients? Did any patients decline to participate? Why? What was the distribution of patients recruited from primary care versus cardiology?
7. More specific details are warranted for the methods used for qualitative analysis, such as whether an inductive versus a deductive design was used. Was a consensus approach used, or some other approach? See also the writing guidelines for qualitative studies (eg, the Consolidated Criteria for Reporting Qualitative Research [COREQ], Standards for Reporting Qualitative Research [SRQR]). Explain also the "additional verification" process during analysis. References should be cited for the qualitative methods used in this work.
8. Did any of the patients have prior experience with MyChart, and if so, what was the average number of years of MyChart experience?
9. These statements from the text appear to be contradictory, and the meaning of the first statement especially is unclear, and seems like an opinion: "[Patients expressed that the] widget should not ask patients to discuss the information with their healthcare provider." "Patients wanted to discuss the recall with their clinicians to 'close the loop.'"

1. The conclusion not to deploy the system seems dramatic based on the findings and makes me wonder if any other creative solutions were considered to address the concern of potential increased clinic burden. Also, how was it determined that the clinic burden outweighed safety risks to the patient? Maybe the system should only be used for certain types of recalls, for example. Or maybe the system could be integrated more with the pharmacy, rather than the prescriber's clinic, or the letter could read differently (advising against contacting the clinic unless the patient was unable to resolve the issue with the pharmacy). Or the letter could explain that only the pharmacy, not the clinic, would have a record of the patient's specific manufacturer and whether the recall applied to them.
2. It would be helpful to see the full interview guide and patient scenario details in a supplementary appendix to aid interpretation of the methods and results.

Discussion

1. The Discussion does not mention limitations of the study design and methods.
2. I expected at least some comparison to other, related literature.
3. Is anything stamped on the medication (eg, pill) itself to indicate the manufacturer? Or is that also inconsistent across medications?
4. A table of key recommendations could strengthen the paper.
5. In the last paragraph of the Discussion, there is no citation for the number of state boards of pharmacy that require the lot number to appear on the label.
6. I expected the Discussion to close with a Conclusions paragraph outlining key lessons learned and any generalizable findings.

Round 2 Review

General Comments

The authors addressed a few of my review comments and made some text changes, but unfortunately, most of my comments—about 15 of them—remain inadequately addressed. For the comments listed again below, the authors did not appear to change anything in the manuscript to address the comment. In many cases, even the authors' reply to the reviewers did not answer the question. Also, the authors describe adding the interview guide as an appendix, but I could not find this file on the reviewer website.

Unaddressed or inadequately addressed review comments are described in the following sections.

Specific Comments

Abstract

1. The Background section appears to be contradictory. Sentence 2 says the Food and Drug Administration has ways to notify HCPs and patients, but then the following sentences seem to say the opposite.

3. The choice of methods doesn't seem to follow the Background section. Why was it necessary to include the clinics, rather than just work directly with the patients? Or, why was the focus on clinics, rather than pharmacies? (These comments apply to the main Introduction and Methods sections, as well.)

Introduction

2. No information is included on if/what literature explores this or similar topics. (Lack of literature citations/review.)

Setting

2. What was the goal sample size and rationale for the sample size? There is missing demographic information on the participating patients.

3. So the FHIR portion notified HCPs? The intended recipients are not specified for that part of the program.

6. Even for a convenience sample, more details are needed on recruitment. How did you choose which patients to email? How many were emailed for recruitment? Were patients emailed and

recruited sequentially, for example? Were there any exclusion or inclusion criteria for patients? Did any patients decline to participate? Why? What was the distribution of patients recruited from primary care versus cardiology?

7. More specific details are warranted for the methods used for qualitative analysis, such as whether an inductive versus a deductive design was used. Was a consensus approach used, or some other approach? See also the writing guidelines for qualitative studies (eg, the COREQ, SRQR). Explain also the "additional verification" process during analysis. References should be cited for the qualitative methods used in this work.

8. Did any of the patients have prior experience with MyChart, and if so, what was the average number of years of MyChart experience?

9. These statements from the text appear to be contradictory, and the meaning of the first statement especially is unclear, and seems like an opinion: "[Patients expressed that the] widget should not ask patients to discuss the information with their healthcare provider." "Patients wanted to discuss the recall with their clinicians to 'close the loop.'"

10. The conclusion not to deploy the system seems dramatic based on the findings and makes me wonder if any other creative solutions were considered to address the concern of potential increased clinic burden. Also, how was it determined that the clinic burden outweighed safety risks to the patient? Maybe the system should only be used for certain types of recalls, for example. Or maybe the system could be integrated more with the pharmacy, rather than the prescriber's clinic, or the letter could read differently (advising against contacting the clinic unless the patient was unable to resolve the issue with the pharmacy). Or the letter could explain that only the pharmacy, not the clinic, would have a record of the patient's specific manufacturer and whether the recall applied to them.

Discussion

1. The Discussion does not mention limitations of the study design and methods.

2. I expected at least some comparison to other, related literature.

3. Is anything stamped on the medication (eg, pill) itself to indicate the manufacturer? Or is that also inconsistent across medications?

4. A table of key recommendations could strengthen the paper.

5. In the last paragraph of discussion, there is no citation for the number of state boards of pharmacy the require the lot number to appear on the label. (The statement that needs a literature citation is "Only three State Boards of Pharmacy require the NDC to appear on the dispensed medication label, and only five State Boards of Pharmacy require the lot number to appear on the dispensed medication label.")

Conflicts of Interest

None declared.

Reference

1. Gadgil M, Pavlakos R, Carini S, et al. Automating individualized notification of drug recalls to patients: complex challenges and qualitative evaluation. JMIRx Med 2026;7:e68345. [doi: 10.2196/68345]

Abbreviations

COREQ: Consolidated Criteria for Reporting Qualitative Research
EHR: electronic health record
FHIR: Fast Healthcare Interoperability Resources
HCP: health care professional
SRQR: Standards for Reporting Qualitative Research

Edited by CN Hang; submitted 18.Aug.2025; this is a non–peer-reviewed article;accepted 18.Aug.2025; published 13.Jan.2026.

Please cite as:
Russ A
Peer Review of "Automating Individualized Notification of Drug Recalls to Patients: Complex Challenges and Qualitative Evaluation"
JMIRx Med 2026;7:e82613
URL: https://xmed.jmir.org/2026/1/e82613
doi:10.2196/82613

Peer Review of “Interpreting the Estimand Framework From a Causal Inference Perspective”

Hao Wu

Michigan State University, East Lansing, MI, United States

Related Articles:

Companion article: http://preprints.jmir.org/preprint/88813

Companion article: https://med.jmirx.org/2026/1/e98121

Companion article: https://med.jmirx.org/2026/1/e88813

(JMIRx Med 2026;7:e98126) doi:10.2196/98126

KEYWORDS

causal inference; clinical trial; estimand; intercurrent event; treatment effect

This is a peer review report for “Interpreting the Estimand Framework From a Causal Inference Perspective.”

Round 1 Review

This manuscript [1] provides a pedagogical interpretation of the International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use (ICH) E9 framework through the lens of the potential outcomes causal inference framework. The author translates the 5 attributes and intercurrent event strategies into formal statistical notation, primarily using simple randomized trial settings and linear models. The paper is largely conceptual and expository in nature, aiming to improve methodological clarity rather than introduce new causal methodology.

Within the context of *JMIRx Med* as an overlay journal for preprints, the manuscript is generally coherent, technically correct at a basic level, and suitable as an educational or perspective-style contribution, though it would likely fall short of expectations for novelty, depth, or rigor in a specialist statistics or causal inference journal.

Major Comments

1. The manuscript repeatedly states that it “interprets” the ICH E9 framework, but in practice, it mostly rephrases ICH E9 concepts using potential outcomes notation. Readers would more likely expect to see discussions on limitations, ambiguities, or contested aspects.
2. While pedagogical simplicity may be intentional, several aspects risk being misleading if read uncritically. For example, conditioning on posttreatment variables (section 3.6) is introduced without adequate warning about collider bias or causal ordering issues, and the discussion of principal stratification glosses over identification challenges, relying on brief mentions of Bayesian methods without clarifying assumptions. These are not fatal flaws, but the author should be more explicit about what is heuristic versus formally justified.

Conflicts of Interest

None declared.

Reference

1. Zeng J. Interpreting the estimand framework from a causal inference perspective. JMIRx Med 2026;7:e88813. [doi: 10.2196/88813]

Abbreviations

ICH: International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use

Edited by A Schwartz; submitted 13.Apr.2026; this is a non–peer-reviewed article;accepted 13.Apr.2026; published 22.May.2026.

Please cite as:
Wu H
Peer Review of "Interpreting the Estimand Framework From a Causal Inference Perspective"
JMIRx Med 2026;7:e98126
URL: https://xmed.jmir.org/2026/1/e98126
doi:10.2196/98126

Peer Review of “Interpreting the Estimand Framework From a Causal Inference Perspective”

Qi Zhang

Emory University School of Medicine, Atlanta, GA, United States

Related Articles:

Companion article: http://preprints.jmir.org/preprint/88813

Companion article: https://med.jmirx.org/2026/1/e98121

Companion article: https://med.jmirx.org/2026/1/e88813

(JMIRx Med 2026;7:e98125) doi:10.2196/98125

KEYWORDS

causal inference; clinical trial; estimand; intercurrent event; treatment effect

This is a peer review report for “Interpreting the Estimand Framework From a Causal Inference Perspective.”

Round 1 Review

General Comments

This manuscript [1] aims to connect the International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use (ICH) E9 (R1) estimand framework with causal inference concepts; however, the exposition suffers from imprecise causal language and several conceptual inaccuracies. In particular, some core causal notions—such as potential outcomes, confounding in randomized trials, and the definition of the average treatment effect (ATE)—are stated incorrectly or inconsistently. In addition, established intercurrent-event strategies from ICH E9 (R1) are presented in a way that may suggest methodological novelty, despite being well known. Clarifying the causal assumptions, correcting the technical definitions, and more clearly distinguishing estimands from estimation methods would substantially improve the rigor and clarity of the manuscript.

Specific Comments

Major Comments

1. The introduction may give the impression that these strategies are newly proposed by the author, whereas they are in fact defined in ICH E9 (R1). The manuscript would benefit from clearer attribution to, and positioning relative to, the ICH E9(R1) estimand framework.

2. The important concept of intercurrent events is not clearly defined. The definition provided in the manuscript, “Intercurrent events are events that happen after treatment initiation and affect the definition of a treatment effect” (page 2) is vague and potentially misleading. It misses the key idea that intercurrent events are posttreatment events that interfere with the interpretation or existence of the outcome relative to the treatment of interest, rather than merely events that affect treatment effects.

3. In section 2, it is incorrect to state that “Ri, Xi and Yi are potential outcomes.” Only Xi (·) and Yi (·) are potential outcomes. The randomization indicator Ri not a potential outcome; it is a realized random variable determined by the design.

4. At the beginning of section 2, the authors assume “an ideal two-arm randomized controlled clinical trial, with full compliance to treatment and no intercurrent events.” In such a setting, confounders do not affect treatment assignment. However, the manuscript later defines “some confounders C that affect both X and Y,” which contradicts the assumption of randomization.

5. ATE is defined as ATE = E(Y (X(R = 1) = 1) | C) − E(Y (X(R = 0) = 0) | C). However, this is a conditional ATE rather than the marginal ATE, since Y (X(R = 1) = 1) and Y (X(R = 0) = 0) are potential outcomes. The author should define the ATE marginally and then mention conditioning on C for adjustment.

6. Page 3: the phrase “the difference (D) between the average treatment effect from participants who take the experimental treatment...” is incorrect. The quantity described is the difference in average observed outcomes, not an average treatment effect.

7. Across strategies, the author repeatedly claims that the estimand formula is “still” the same, which is misleading. The symbolic form may look similar, but the estimand is not the same. In treatment policy, X is redefined; in composite and while-on-treatment strategies, Y is redefined; in principal stratification, the target population changes. This undermines the central E9 (R1) message that different strategies define different estimands.

8. The proposed “model adjustment strategy” does not correspond to an estimand strategy as defined in ICH E9 (R1), but rather to a particular modeling or estimation approach.

Moreover, in case 1 of Figure 2, concomitant therapies occur after treatment initiation, which is inconsistent with the causal diagram in Figure 3. In this setting, M may act as a mediator rather than a confounder. Treating postrandomization intercurrent events as confounders requires careful causal justification and may induce bias; this issue is not discussed in the manuscript.

Minor Comments

9. Please spell out the abbreviation "ICH" at its first occurrence.

10. Some sentences are confusing and would benefit from revision. For example, in the second paragraph on page 2: "Intercurrent events are frequent in practice but conceptually novel. E9(R1) listed many examples for intercurrent events, such as use of concomitant therapies, treatment switching and death before endpoint measurement." Intercurrent events are not really new conceptually; rather, they were newly formalized or explicitly emphasized in E9 (R1). In "examples for intercurrent events," the preposition should be "of," not "for." As a second example, "This individual treatment effect controls confounders on the endpoint within the same participant and means how the endpoint would change when only the treatment condition changes" on page 3: the ITE does not "control confounders"; it is defined counterfactually for the same individual.

11. A right parenthesis is missing in the first ATE formula on page 3.

12. Equation 2.1 is missing the observed randomization indicator R^o in the first line.

13. The exclusion restriction assumption for instrumental variables should be stated more clearly.

Conflicts of Interest

None declared.

Reference

1. Zeng J. Interpreting the estimand framework from a causal inference perspective. JMIRx Med 2026;7:e88813. [doi: 10.2196/88813]

Abbreviations

ATE: average treatment effect
ICH: International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use

Edited by A Schwartz; submitted 13.Apr.2026; this is a non–peer-reviewed article;accepted 13.Apr.2026; published 22.May.2026.

Please cite as:
Zhang Q
Peer Review of "Interpreting the Estimand Framework From a Causal Inference Perspective"
JMIRx Med 2026;7:e98125
URL: https://xmed.jmir.org/2026/1/e98125
doi:10.2196/98125

Peer Review of “Interpreting the Estimand Framework From a Causal Inference Perspective”

Linying Zhang

Washington University in St Louis, St. Louis, MO, United States

Related Articles:

Companion article: http://preprints.jmir.org/preprint/88813

Companion article: https://med.jmirx.org/2026/1/e98121

Companion article: https://med.jmirx.org/2026/1/e88813

(JMIRx Med 2026;7:e98122) doi:10.2196/98122

KEYWORDS

causal inference; clinical trial; estimand; intercurrent event; treatment effect

This is a peer review report for “Interpreting the Estimand Framework From a Causal Inference Perspective.”

Round 1 Review

This Viewpoint [1] uses mathematical expressions to formalize the estimand framework proposed by a professional society (ie, the ICH), targeting a pharmaceutical industry audience involved in clinical trial design and analysis. While the mathematical formulations for the estimation strategies appear technically correct, the article lacks clarity regarding the significance of the problem it aims to address. The overall discussion remains superficial, offering limited conceptual or practical contributions to the existing literature.

Major Comments

1. The professional society “ICH” is never spelled out or introduced. Additional context is needed regarding the role of the ICH, its influence on regulatory science, and why its guidelines are particularly important for clinical trial design and analysis.
2. The Efficacy Guideline E9 was published in 2019. The authors should clarify what impact this guideline has had on the pharmaceutical industry since its release. Moreover, it is unclear why a causal interpretation of this guideline is timely and important in 2025, several years after its publication.
3. None of the proposed strategies address noncompliance, such as cases where treatment is not received despite assignment or is received without assignment (eg, X(R=1)=0 or X(R=0)=1). Noncompliance is a central issue in causal inference and should be explicitly discussed. If noncompliance is assumed to be irrelevant, then the introduction of the notation R appears redundant and should be justified or removed.
4. The strategies are presented at a very high level. Although the 4 cases illustrated in Figure 2 provide some intuition regarding the appropriateness of each strategy, the Viewpoint would be substantially strengthened by grounding the discussion in real clinical trial examples. Demonstrating how each strategy has been applied in practice would greatly improve clarity and impact.
5. The scope and framing of the Viewpoint appear better suited for a pharmaceutical science or regulatory-focused journal rather than a JMIR-based journal. The authors should better justify the relevance of this work to the JMIR readership or reconsider the target venue.

Minor Comments

1. Section 2 begins with the statement: “A causal inference framework is based on the potential outcome framework.” This is inaccurate, as causal inference can also be grounded in other frameworks, such as structural causal models.
2. In the abstract, the sentence “This article aims to interpret the estimand framework through its underlying theories, the causal inference framework based on potential outcomes” should replace the comma with “and” for grammatical correctness.
3. On page 2, second line: “Generally, Treaments are…” — the “T” in “Treatments” should not be capitalized.
4. In section 3.2 (page 6): the sentence “Through the hypothetical strategy, we make the second…” is ambiguous and should be rewritten for clarity.

Round 2 Review

I appreciate the authors taking the time to address the reviewers’ comments. I have read the revised manuscript, but I still find that the Viewpoint is too incremental and not a strong fit for the journal’s audience. It is not sufficiently well motivated why translating this particular guideline into the potential outcomes framework is important for researchers across settings, from clinical trials to observational studies. While the potential outcomes framework is already widely used, it remains unclear

why aligning it specifically with this guideline represents a meaningful or novel contribution that warrants a Viewpoint article.

Conflicts of Interest

None declared.

Reference

1. Zeng J. Interpreting the estimand framework from a causal inference perspective. JMIRx Med 2026;7:e88813. [doi: 10.2196/88813]

Edited by A Schwartz; submitted 13.Apr.2026; this is a non–peer-reviewed article;accepted 13.Apr.2026; published 22.May.2026.

Please cite as:
Zhang L
Peer Review of "Interpreting the Estimand Framework From a Causal Inference Perspective"
JMIRx Med 2026;7:e98122
URL: https://xmed.jmir.org/2026/1/e98122
doi:10.2196/98122

Peer Review of "The Performance of DeepSeek R1 and Gemini 3 in Complex Medical Scenarios: Comparative Study"

Ziyu Wang

University of California, Irvine, Irvine, CA, United States

Related Articles:

Companion article: https://www.medrxiv.org/content/10.1101/2025.04.29.25326666v1

Companion article: https://med.jmirx.org/2026/1/e96220

Companion article: https://med.jmirx.org/2026/1/e76822

(JMIRx Med 2026;7:e96223) doi:10.2196/96223

KEYWORDS

large reasoning model; LRM; large language model; LLM; accuracy; medical scenario; DeepSeek R1; Gemini 3

This is a peer-review report for "The Performance of DeepSeek R1 and Gemini 3 in Complex Medical Scenarios: Comparative Study."

Round 1 Review

General Comments

This is a timely and well-structured paper [1] that investigates the application of DeepSeek R1, a state-of-the-art large reasoning model, in the medical domain using the Multitask Language Understanding Pro (MMLU-Pro) benchmark. This paper presents a follow-up evaluation of the DeepSeek R1 large reasoning model on open-ended medical scenarios from the MMLU-Pro benchmark. The study finds that DeepSeek R1 achieves a high accuracy of 92% without multiple-choice options, demonstrating its potential utility in more realistic clinical settings. The paper is timely and relevant, with strong empirical results and clear motivation. However, it would benefit from revisions to improve clarity, contextual grounding in existing work, and methodological detail. The authors may also consider citing recent work that examines the questioning strategies of large language models (LLMs) in clinical dialogues to better position this study in the broader landscape.

The study is commendable for its effort in combining expert validation with benchmark testing and highlighting both performance and interpretability aspects. The paper is generally well-written, informative, and relevant to the research community on artificial intelligence (AI) in health care.

However, before being suitable for publication, several important revisions are required. These include expanding the related work section to better situate the contribution of current research efforts, addressing some methodological limitations more transparently, and improving the robustness and generalizability of conclusions. Thus, I recommend revision and re-review.

Specific Comments

Major Comments

1. While the paper references MMLU, MedQA, and some domain-specific LLM evaluations, it lacks a deeper discussion on recent approaches to questioning capabilities and long-context understanding in medical AI. Two notable papers should be included. First, "HealthQ: Unveiling Questioning Capabilities of LLM Chains in Healthcare Conversations" by Wang et al [2]. This paper presents a benchmarking framework focusing on the inquiry and elicitation capacity of LLM chains, which directly relates to the "reasoning" and prompt design aspects discussed here. Second, "Context Clues: Evaluating Long Context Models for Clinical Prediction Tasks on EHR Data" by Wornow et al [3]. This study highlights how context windows and task framing affect LLM performance on clinical reasoning—relevant for understanding how question complexity and format might interact with LLM accuracy.

The paper could be strengthened by referencing more recent work on prompting and questioning strategies in clinical LLM applications. The paper would benefit from referencing Wang et al [2], which evaluates LLM chains' ability to optimize questions through reflection and prompting. This is relevant to the current paper's interest in open-ended diagnostic reasoning and LLM behavior in clinical settings.

2. Although the DeepSeek R1 model is rigorously evaluated against MMLU-Pro, there's a lack of direct performance comparison to other LLMs (eg, MedPaLM, GPT-4, Claude) on the same dataset or medical scenarios. Even informal or partial benchmarks would help contextualize the model's effectiveness. Also, the novelty should be better emphasized—is this the first comprehensive large reasoning model evaluation on MMLU-Pro's health subset?

3. The paper rightly points out issues with cueing and "testwiseness" in multiple-choice questions but doesn't propose concrete mitigations. The planned future work of testing without

answer choices is excellent—consider incorporating a small pilot of this now or discussing expected outcomes in more depth. Also, the limitations of using only 162 scenarios across many specialties could be made more transparent, especially regarding statistical robustness and specialty-specific insights.

4. The study uses a fixed prompt but does not explore or discuss the impact of prompt variations, which may influence results in open-ended tasks.

5. While the discussion of biases and failure modes is helpful, a more structured breakdown of error types and their frequency would improve the interpretability of findings.

6. The discussion on reasoning steps and transparency is insightful but could be expanded to address recent concerns about the faithfulness of chain-of-thought outputs.

Minor Comments

7. Model latency and usability: while the latency of DeepSeek is acknowledged, it's not contextualized with respect to potential clinical utility or workflow integration. A brief paragraph on practical deployment implications would strengthen the discussion.

8. Citation formatting: ensure all references (especially web-based ones like Perplexity and PromptHub) are consistently formatted and maintained in the reference list.

9. Future directions could be made more actionable by suggesting benchmark expansions with real patient data or multimodal inputs.

Round 2 Review

My comments have been addressed.

Conflicts of Interest

None declared.

References

1. Bajwa M, Hoyt R, Knight D, Haider M. The performance of DeepSeek R1 and Gemini 3 in complex medical scenarios: comparative study. JMIRx Med 2026;7:e76822. [doi: 10.2196/76822]
2. Wang Z, Li H, Huang D, Kim HS, Shin CW, Rahmani AM. HealthQ: unveiling questioning capabilities of LLM chains in healthcare conversations. Smart Health (2014) 2025 Jun;36:100570. [doi: 10.1016/j.smhl.2025.100570]
3. Wornow M, Bedi S, Hernandez MAF, et al. Context clues: evaluating long context models for clinical prediction tasks on EHR data. Presented at: ICLR 2025; the Thirteenth International Conference on Learning Representations; Apr 24-28, 2025.

Abbreviations

AI: artificial intelligence
LLM: large language model
MMLU-Pro: Multitask Language Understanding Pro

Edited by A Schwartz; submitted 26.Mar.2026; this is a non–peer-reviewed article;accepted 26.Mar.2026; published 27.Apr.2026.

Please cite as:
Wang Z
Peer Review of "The Performance of DeepSeek R1 and Gemini 3 in Complex Medical Scenarios: Comparative Study"
JMIRx Med 2026;7:e96223
URL: https://xmed.jmir.org/2026/1/e96223
doi:10.2196/96223

Peer Review of "The Performance of DeepSeek R1 and Gemini 3 in Complex Medical Scenarios: Comparative Study"

Jacqueline Guan-Ting You

Mass General Brigham, Boston, MA, United States

Related Articles:

Companion article: https://www.medrxiv.org/content/10.1101/2025.04.29.25326666v1

Companion article: https://med.jmirx.org/2026/1/e96220

Companion article: https://med.jmirx.org/2026/1/e76822

(JMIRx Med 2026;7:e96225) doi:10.2196/96225

KEYWORDS

large reasoning model; LRM; large language model; LLM; accuracy; medical scenario; DeepSeek R1; Gemini 3

This is a peer-review report for "The Performance of DeepSeek R1 and Gemini 3 in Complex Medical Scenarios: Comparative Study."

Round 1 Review

General Comments

This paper [1] seeks to evaluate the accuracy of DeepSeek R1 in correctly identifying the primary medical diagnosis in the medical scenarios dataset portion of Massive Multitask Language Understanding Pro (MMLU-Pro) using an open-ended format. Some clarifications on the methods and results (especially around the roles of subject matter experts vs core team members in the publication), would be helpful in understanding how these results were derived.

Specific Comments

Minor Comments

1. Introduction: consider citing Deepseek AI's Deepseek R1 paper [2].
2. Methods: please clarify who your subject matter experts were (eg, physicians, researchers) in terms of rank, specialty, and role and how they were used to grade answers (eg, selected based on specialty, 2 reviewer process, etc).
3. Methods: please indicate when the analyses were run.
4. Results: who determines whether references are related or unrelated?
5. Results and Discussion: it is unclear to me from reading the discussion portion of the paper as to whether we have any sense of whether DeepSeek R1 has correct reasoning for questions with correct diagnoses (eg, it may get the right diagnosis but may have incorrect reasoning). Similarly, did you determine the "correct answer" based on string matching (for example, if the answer was "septic arthritis" and the DeepSeek output stated "septic shock," would this be incorrect)?
6. Discussion: consider acknowledging the sample size of questions as a limitation.

Round 2 Review

General Comments

The paper has been revised to address the Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis–Large Language Model (TRIPOD-LLM) guidelines. Overall, it appears most concerns from both reviewers have been addressed.

Conflicts of Interest

None declared.

References

1. Bajwa M, Hoyt R, Knight D, Haider M. The performance of DeepSeek R1 and Gemini 3 in complex medical scenarios: comparative study. JMIRx Med 2026;7:e76822. [doi: 10.2196/76822]
2. Guo D, Yang D, Zhang H, et al. DeepSeek-R1: incentivizing reasoning capability in LLMs via reinforcement learning. . Preprint posted online on Jan 22, 2025. [doi: 10.48550/arXiv.2501.12948]

Abbreviations

MMLU-Pro: Massive Multitask Language Understanding Pro
TRIPOD-LLM: Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis–Large Language Model

Edited by A Schwartz; submitted 26.Mar.2026; this is a non–peer-reviewed article;accepted 26.Mar.2026; published 27.Apr.2026.

Please cite as:
You JGT
Peer Review of "The Performance of DeepSeek R1 and Gemini 3 in Complex Medical Scenarios: Comparative Study"
JMIRx Med 2026;7:e96225
URL: https://xmed.jmir.org/2026/1/e96225
doi:10.2196/96225

Peer Review of “The Performance of DeepSeek R1 and Gemini 3 in Complex Medical Scenarios: Comparative Study”

Mayank Kejriwal

University of Southern California, Los Angeles, CA, United States

Related Articles:

Companion article: https://www.medrxiv.org/content/10.1101/2025.04.29.25326666v1

Companion article: https://med.jmirx.org/2026/1/e96220

Companion article: https://med.jmirx.org/2026/1/e76822

(JMIRx Med 2026;7:e96227) doi:10.2196/96227

KEYWORDS

large reasoning model; LRM; large language model; LLM; accuracy; medical scenario; DeepSeek R1; Gemini 3

This is a peer-review report for “The Performance of DeepSeek R1 and Gemini 3 in Complex Medical Scenarios: Comparative Study.”

Round 1 Review

General Comments

This paper [1] reports on an experimental study to analyze the Massive Multitask Language Understanding Pro (MMLU-Pro) Q&A dataset. The authors find that DeepSeek R1 had an accuracy rate of 95.1% in 162 medical scenarios after reconciliation with subject matter experts on 23 questions. The findings contribute to the growing body of knowledge on large language model applications in health care and provide insights into the strengths and limitations of DeepSeek R1 in this domain.

Specific Comments

Major Comments

1. The results are not appropriately qualified with results on statistical significance, and/or are lacking comparisons with other language models. Even if we know how other models perform overall, it would still be good to have more details, such as a comparison of where one model is right and another is wrong. Those kinds of deep insights are lacking in this paper. All we really know is that DeepSeek performs at a level roughly equivalent to the other leading models (nothing surprising there) and that it sometimes has incomplete or inexplicable behavior. I feel the paper needs to have more results and analysis to be a good fit for this journal.

2. Maybe you could add a workflow diagram/figure to better illustrate the methods?

3. I would like Table 1 to be augmented. Perhaps you can add an example question with answer choices? Right now, it looks very trivial. The alternative is to create a simple bar graph instead of a table, but the former would be more useful.

Conflicts of Interest

None declared.

Reference

1. Bajwa M, Hoyt R, Knight D, Haider M. The performance of DeepSeek R1 and Gemini 3 in complex medical scenarios: comparative study. JMIRx Med 2026;7:e76822. [doi: 10.2196/76822]

Abbreviations

MMLU-Pro: Massive Multitask Language Understanding Pro

Edited by A Schwartz; submitted 26.Mar.2026; this is a non–peer-reviewed article;accepted 26.Mar.2026; published 27.Apr.2026.

Please cite as:
Kejriwal M
Peer Review of "The Performance of DeepSeek R1 and Gemini 3 in Complex Medical Scenarios: Comparative Study"
JMIRx Med 2026;7:e96227
URL: https://xmed.jmir.org/2026/1/e96227
doi:10.2196/96227

Peer Review of “Awareness, Experiences, and Attitudes Toward Preprints Among Medical Academics: Convergent Mixed Methods Study”

Ludo Waltman

Leiden University, Leiden, The Netherlands

Related Articles:

Companion article: https://www.medrxiv.org/content/10.1101/2025.04.09.25325524v1

Companion article: https://med.jmirx.org/2026/1/e95735

Companion article: https://med.jmirx.org/2026/1/e78139

(JMIRx Med 2026;7:e95737) doi:10.2196/95737

KEYWORDS

preprint; medical academics; publishing attitudes; editorial policies; survey

This is a peer review report for “Awareness, Experiences, and Attitudes Toward Preprints Among Medical Academics: Convergent Mixed Methods Study.”

Round 1 Review

This paper [1] offers valuable insights into attitudes toward preprinting at a medical institution in Istanbul. The research appears sound to me. The paper is clear and easy to read.

I have one important comment: I was unable to find the survey form. I urge the authors to make the survey form, and also the data collected in the survey, openly available. To interpret the results of the survey, it is important to have access to the underlying survey questions.

Round 2 Review

In their response to my earlier comments and the comments of the other reviewer, the authors refer to supplementary material. I am unable to find this supplementary material. Where can I find it?

Apart from this issue, I am satisfied with the revised article. I have no suggestions for further improvements.

Conflicts of Interest

None declared.

Reference

1. Sevim M, Karamese B, Alparslan Z. Awareness, experiences, and attitudes toward preprints among medical academics: convergent mixed methods study. JMIRx Med 2025;7:e78139. [doi: 10.2196/78139]

Edited by S Amal; submitted 19.Mar.2026; this is a non–peer-reviewed article;accepted 19.Mar.2026; published 17.Apr.2026.

Please cite as:
Waltman L
Peer Review of “Awareness, Experiences, and Attitudes Toward Preprints Among Medical Academics: Convergent Mixed Methods Study”
JMIRx Med 2026;7:e95737
URL: https://xmed.jmir.org/2026/1/e95737
doi:10.2196/95737

Peer Review of “Awareness, Experiences, and Attitudes Toward Preprints Among Medical Academics: Convergent Mixed Methods Study”

Kazuki Ide

Osaka University, Osaka, Japan

Related Articles:

Companion article: https://www.medrxiv.org/content/10.1101/2025.04.09.25325524v1

Companion article: https://med.jmirx.org/2026/1/e95735

Companion article: https://med.jmirx.org/2026/1/e78139

(JMIRx Med 2026;7:e95736) doi:10.2196/95736

KEYWORDS

preprint; medical academics; publishing attitudes; editorial policies; survey

This is a peer review report for “Awareness, Experiences, and Attitudes Toward Preprints Among Medical Academics: Convergent Mixed Methods Study.”

Round 1 Review

General Comments

This paper [1] reports on a survey exploring the awareness, experiences, and attitudes toward preprints among medical faculty at a university in Türkiye. While the topic may be of some relevance to the academic community, the manuscript contains fundamental issues that must be addressed. As it currently stands, the manuscript is not suitable for publication in the journal. Specific comments are provided below.

Specific Comments

Major Comments

1. Abstract: The authors should specify the name of the university referred to as “a major university in Istanbul.”
2. Abstract: The authors should clarify how the responding editors and the biomedical journals that were manually reviewed were selected.
3. Abstract: The authors are encouraged to present concrete data rather than relying solely on descriptive summaries.
4. Introduction: The authors describe the study as “mixed method,” but it lacks a legitimate integration process between the qualitative and quantitative components. Therefore, the term “mixed method” should be avoided unless such integration is clearly demonstrated.
5. Methods: The authors should specify the name of the university referred to as “a major medical university in Istanbul.” This information is important for assessing the reliability of the study and for confirming ethical approval in a transparent manner.
6. Methods: The authors should indicate the total number of potential participants (ie, the total number of faculty members invited or eligible to participate).
7. Methods: The authors should explain the rationale for dividing the age groups at 40 years.
8. Methods: The authors are encouraged to classify the biomedical journals into basic and clinical categories, in the same way that they categorized the survey respondents, even if some journals may cover both areas.
9. Methods: The authors should provide a list of the journals included in this study.
10. Methods: The section lacks a description of statistical analysis, making the analytical process unclear and only partially reported.
11. Results: The authors should ensure that the findings are presented in alignment with the methods described. As the study does not involve a systematic review but rather a journal policy review, the current framing of the Results section may give a misleading impression.
12. Results: For clarity and coherence, the Results section should be reorganized to reflect the sequence of the study components—for example, starting with the questionnaire survey results, followed by the findings from the editorial and journal policy survey.
13. Results: The authors mention the impact factor, but it is not described in the Methods section. Furthermore, the year and whether it represents the 2-year or 5-year impact factor are not specified.
14. Results: The description of the preprint test, including its content and scoring method, is insufficient, making it difficult to assess its appropriateness.
15. Results: The process for analyzing the qualitative responses is not clearly described, and the presentation of the results is extremely limited.

16. Results: The authors should provide information on how the responses from the editors were summarized. Without this explanation, reviewers and potential readers may find it difficult to interpret the results presented.
17. Discussion: The authors should revise the manuscript for logical consistency and explicitly discuss the limitations of this study prior to submitting it to another journal.

Minor Comments

1. Tables: The authors should ensure consistent use of commas, periods, and digit formatting. Furthermore, the tables contain typographical errors that need correction.
2. References: The authors should review the reference formatting and ensure that it adheres to the journal's prescribed style.

Conflicts of Interest

None declared.

Reference

1. Sevim M, Karamese B, Alparslan Z. Awareness, experiences, and attitudes toward preprints among medical academics: convergent mixed methods study. JMIRx Med 2025;7:e78139. [doi: 10.2196/78139]

Edited by S Amal; submitted 19.Mar.2026; this is a non–peer-reviewed article;accepted 19.Mar.2026; published 17.Apr.2026.

Please cite as:
Ide K
Peer Review of "Awareness, Experiences, and Attitudes Toward Preprints Among Medical Academics: Convergent Mixed Methods Study"
JMIRx Med 2026;7:e95736
URL: https://xmed.jmir.org/2026/1/e95736
doi:10.2196/95736

Peer Review of “Associations Between IT Job Stressors and Anxiety, Depression, and Stress: Cross-Sectional Study”

Holger Mühlan

Universität Greifswald, Greifswald, Germany

Related Articles:

Companion article: https://www.medrxiv.org/content/10.1101/2025.02.24.25322785v1

Companion article: https://med.jmirx.org/2026/1/e91437

Companion article: https://med.jmirx.org/2026/1/e73211

(JMIRx Med 2026;7:e91383) doi:10.2196/91383

KEYWORDS

survey; association; occupational health; mental health; stressors; IT; IT professionals; United States; workplace; depression; anxiety; stress; help-seeking; health literacy

This is a peer review report for “Associations Between IT Job Stressors and Anxiety, Depression, and Stress: Cross-Sectional Study.”

Round 1 Review

General Comments

Although the data basis of this study [1] is not very reliable due to methodological limitations (cross-sectional design, online survey, self-reporting only, self-selection bias), the study does provide some interesting insights. I have the following comments to make.

Specific Comments

Major Comments

What I miss most is a more differentiated discussion of the assumed mediation by mental health literacy (MHL). In my view, it is not necessarily plausible. It would be just as plausible to assume that people with high MHL are more competent in knowing what to do by themselves. In addition, MHL itself can also have a protective function, in that people with high MHL also know what they can do by themselves to protect themselves in the event of a high workload and apply any compensatory measures (eg, relaxation).

Minor Comments

1. Some information on the analysis should be added to the abstract.
2. “Anxiety” should also be included in the keywords.
3. With regard to the excluded cases, no cases are mentioned that were excluded due to conspicuous response behavior (eg, monotonous patterns) or too rapid completion (“speeders”). Why?
4. Please add information on how many people in total were initially contacted.
5. A core element of the study is the initial identification of IT-specific stressors. Here, further information on the criteria for the selection of the experts and the methodological procedure for capturing the stressors is essential (interviews? focus group? workshop?).
6. The study just assessed the intention to seek help. Did it also assess whether professional help had been sought in the past 12 months? If not, why not? This would have been very easy to capture and a much more reliable criterion than just intentions.
7. Please add a table with the most important information about the sample.
8. Please also add a table with the frequencies of the individual stressors as well as the distribution of multiple stressors (ie, how often people reported 1 stressor, 2 stressors, etc, up to 12 stressors). This is also interesting, as the effect of multiple stressors appears to be surprisingly small. The type of stressor therefore seems to be more decisive than the frequency/diversity.
9. Table 1 has a different font type, please adjust.
10. Some of the terms in Table 2 are written inconsistently (eg, “p-value”/“P value”).
11. Please add a legend beneath Table 2, explaining the abbreviations of the measures.
12. The discussion basically only addresses why mediation shows no effect with regard to stress, but not why this is also the case with anxiety. Please also address this.
13. Incidentally, the mediation effect for depression should not be overestimated. The effect just tips significance and the size of the indirect effect is rather small relative to the huge direct effect.
14. The assessment of MHL by self-report is not necessarily a limitation per se, as there are both objective and subjective concepts in MHL. The question is rather which version is the more suitable for operationalization for testing your hypotheses.

15. The references are still very inconsistent, eg, some journal titles are abbreviated/others are not; some references lack information (eg, Northwave—Where did “In” appear?) or the year of publication (eg, for Boehm et al [2]). Please check the reference list manually throughout.

Round 2 Review

The authors addressed all points appropriately and I have not spotted any further issues.

Conflicts of Interest

None declared.

References

1. Garcia Colato E, Liu N, Chow A, Sherwood-Laughlin CM, Macy JT. Associations between IT job stressors and anxiety, depression, and stress: cross-sectional study. JMIRx Med 2026;7:e73211. [doi: 10.2196/73211]
2. Boehm MA, Lei QM, Lloyd RM, Prichard JR. Depression, anxiety, and tobacco use: overlapping impediments to sleep in a national sample of college students. J Am Coll Health 2016 Oct;64(7):565-574. [doi: 10.1080/07448481.2016.1205073] [Medline: 27347758]

Abbreviations

MHL: mental health literacy

Edited by A Grover; submitted 13.Jan.2026; this is a non–peer-reviewed article;accepted 13.Jan.2026; published 03.Mar.2026.

Please cite as:
Mühlan H
Peer Review of “Associations Between IT Job Stressors and Anxiety, Depression, and Stress: Cross-Sectional Study”
JMIRx Med 2026;7:e91383
URL: https://xmed.jmir.org/2026/1/e91383
doi:10.2196/91383

Peer Review of "Investigating the Variable Component of the Systematic Error, a Neglected Error Parameter: Theoretical Reevaluation Study"

Elvar Theodorsson

Linköping University, Linköping, Sweden

Related Articles:

Companion article: https://www.medrxiv.org/content/10.1101/2023.05.24.23290382v1

Companion article: https://med.jmirx.org/2026/1/e88981

Companion article: https://med.jmirx.org/2026/1/e49657

Abstract

(JMIRx Med 2026;7:e88830) doi:10.2196/88830

KEYWORDS

repeatability condition; reproducibility within laboratory condition, measurement; systematic error; clinical laboratory; quality control; bias; QC; statistical; statistics; mathematics; computer simulation; standard deviation

This is the peer-review report for "Investigating the Variable Component of the Systematic Error, a Neglected Error Parameter: Theoretical Reevaluation Study."

Round 1 Review

The phenomenon the author of this paper [1] calls "the variable component of the systematic error" (VCSE) is highly important and relevant, and the approach proposed by the author is worthy of serious scientific dialog. However, a prerequisite for serious dialog is (1) a well-structured manuscript based on (2) extensive knowledge of the state of the art in calculating measurement uncertainty and (3) well-written English text.

Unfortunately, this manuscript fails in all three aspects. As a reviewer, I urge the author to seek collaboration with a scientist even better versed in the actual field(s) than himself to create a more deserving manuscript.

1. The handling of constant or intermittent bias has been a challenge for more than 200 years, especially since Gauss and Laplace's work in the early 19th century. The author refers to Eisenhart's excellent 1963 paper, which is appropriate but not as the origin of VCSE. Shewhart's 1923 and 1939 books [2,3] also address this matter.

2. Bias, including VCSE, is also a major point of contention in the International Bureau of Weights and Measures/International Organization for Standardization work on the Guide to the Expression of Uncertainty in Measurement and its revision. The complexity of the issues is illustrated, for example, in a book by Krystek [4]. The current manuscript illustrates the opinions of its author but fails to illustrate the background of the immense scientific literature and debates that have already dealt with the matter.

3. The author needs to clarify whether he adheres to the error or uncertainty paradigms in measurement uncertainty/error questions. The current manuscript represents a mixture of both.

4. The "thought-provoking and even shocking" fact that the Westgard rules and power calculations are based on repeatability uncertainty and not on reproducibility uncertainty is well-known by metrologists in clinical chemistry. Unfortunately, mentioning this fact commonly hurts the sentiments of a majority in our field, and many of us avoid harping on it. A prerequisite for appropriately using Westgard rules and variants is that changing goal mean values are used as calibrators, and reagent lots change over time.

5. The traceability hierarchies used when producing reference materials and calibrators are usually claimed to explain the variations experiences (eg, during lot-number changes). The author apparently does not accept this explananation of the main cause of lot-number shifts/bias, and he needs to explain why his mathematical/statistical theory should be accepted instead.

6. In a crucial part of his manuscript, the author claims that "While RE changes unpredictably from measurement to measurement, VCSE(t) [variable component of systematic error at the moment t] remains quasi-constant in a given day, influencing all measurement results obtained in that day systematically. But in long-term experiments, VCSE(t) becomes a cyclical time-variable function, which repeats the same values after unequal periods. (A period may last even one month)."

The author presents Cobas 6000 analyzer data in support of his thesis. However, data from a variety of measuring systems, lot changes, and measurands are needed before this theory of a cyclical phenomenon is chosen instead of a theory of random components.

7. The author's approach deserves to be published in a better-structured manuscript, written in far better English than the English language of the present manuscript.

Round 2 Review

In the first round of reviews, I asked for "(1) a well-structured manuscript based on (2) extensive knowledge of the state of the art in calculating measurement uncertainty and (3) well-written English text."

The revised version of the manuscript has improved the English text but needs to improve in the two other aspects.

I agree that the paper's subject is essential. The author is well-versed in mathematical statistics and has practical experience in laboratory quality control. However, the manuscript lacks in:

- 1. Counting in metrological aspects
- 2. Using a conventional manuscript structure
- 3. Showing sufficient real laboratory results and the consequence of using the proposed paradigm on real laboratory results

Conflicts of Interest

None declared.

References

1. Vandra AB. Investigating the variable component of the systematic error, a neglected error parameter: theoretical reevaluation study. JMIRx Med 2025;7:e49657. [doi: 10.2196/49657]
2. Shewhart WA. Economic Control of Quality of Manufactured Product: D. Van Nostrand Company; 1923.
3. Shewhart WA. Statistical Method from the Viewpoint of Quality Control: Dover Publications; 1939.
4. Krystek M. Calculating Measurement Uncertainties: Beuth Verlag GmbH; 2016.

Abbreviations

VCSE: variable component of systematic error
VCSE(t): variable component of systematic error at the moment t

Edited by T Leung; submitted 02.Dec.2025; this is a non–peer-reviewed article;accepted 02.Dec.2025; published 27.Feb.2026.

Please cite as:
Theodorsson E
Peer Review of "Investigating the Variable Component of the Systematic Error, a Neglected Error Parameter: Theoretical Reevaluation Study"
JMIRx Med 2026;7:e88830
URL: https://xmed.jmir.org/2026/1/e88830
doi:10.2196/88830

Peer Review of “Investigating the Variable Component of the Systematic Error, a Neglected Error Parameter: Theoretical Reevaluation Study”

Anonymous

Related Articles:

Companion article: https://www.medrxiv.org/content/10.1101/2023.05.24.23290382v1

Companion article: https://med.jmirx.org/2026/1/e88981

Companion article: https://med.jmirx.org/2026/1/e49657

(JMIRx Med 2026;7:e90221) doi:10.2196/90221

KEYWORDS

repeatability condition; reproducibility within laboratory condition, measurement; systematic error; clinical laboratory; quality control; bias; QC; statistical; statistics; mathematics; computer simulation; standard deviation

This is the peer-review report for “Investigating the Variable Component of the Systematic Error, a Neglected Error Parameter: Theoretical Reevaluation Study.”

Round 1 Review

General Comments

With an emphasis on the idea of VCSE(t) (variation in control sample error over time), the study [1] provides a thorough examination of variation in laboratory data. The study emphasizes the significance of differentiating between random and systematic errors, suggests fresh approaches for precisely calculating sVCSE (an SD), and supports updated quality control procedures.

Specific Comments

Major Comments

- 1. The study does not provide empirical confirmation of suggested approaches using real-world data, although it mentions computer simulations and experimental verification. The suggested methodologies’ efficacy and dependability are yet unknown in the absence of empirical validation.
- 2. Linear drifts in daily means across time are assumed in the study. Numerous factors, such as the environment, instrument calibration, and reagent stability, can affect real-world drift patterns and lead to nonlinear trends in daily means over time. The study might have simplified the complicated nature of drift processes by assuming linearity, which could result in estimates of mean values and error components that are not true.
- 3. The assumption that information from internal quality control sources alone can be used to accurately calculate VCSE(t) is inaccurate. Even though internal quality control data offer insightful information on short-term variability, they might not include all sources of variation, particularly those pertaining to outside variables like shifts in the environment, instrument performance, or operator technique. Ignoring these outside influences could result in an inaccurate or understated VCSE(t), which would compromise the validity of the suggested quality control techniques.

Minor Comments

- 1. Although there is a suggestion in the Conclusions section that the present quality control paradigm needs to be revised, there is no concrete plan or set of recommendations based on statistical or mathematical concepts.
- 2. The paper lacks a Discussion section, which could have allowed the author to interpret and contextualize the study’s findings. Additionally, it could have provided an opportunity to compare the study’s findings with previous studies, discuss their implications, and address potential sources of error or bias.

Conflicts of Interest

None declared.

Reference

1. Vandra AB. Investigating the variable component of the systematic error, a neglected error parameter: theoretical reevaluation study. JMIRx Med 2026;7:e49657. [doi: 10.2196/49657]

Abbreviations

VCSE(t): variation in control sample error over time

Edited by T Leung; submitted 23.Dec.2025; this is a non–peer-reviewed article;accepted 23.Dec.2025; published 27.Feb.2026.

Please cite as:
Anonymous
Peer Review of "Investigating the Variable Component of the Systematic Error, a Neglected Error Parameter: Theoretical Reevaluation Study"
JMIRx Med 2026;7:e90221
URL: https://xmed.jmir.org/2026/1/e90221
doi:10.2196/90221

Peer Review of “Automating Individualized Notification of Drug Recalls to Patients: Complex Challenges and Qualitative Evaluation”

Robert Carter Marshall

Madigan Army Medical Center, 2001 28th Street Court NW, Gig Harbor, WA, United States

Related Articles:

Companion article: https://www.medrxiv.org/content/10.1101/2024.09.18.24312141v1

Companion article: https://med.jmirx.org/2026/1/e82609

Companion article: https://med.jmirx.org/2026/1/e68345

(JMIRx Med 2026;7:e82612) doi:10.2196/82612

KEYWORDS

notification system; drug recalls; patient safety; medication; electronic health records; prescriptions; decision support

This is a peer review report for “Automating Individualized Notification of Drug Recalls to Patients: Complex Challenges and Qualitative Evaluation.”

Round 1 Review

General Comments

This paper [1] describes a qualitative study that aims to leverage the US Food and Drug Administration (FDA)’s Healthy Citizen prototype platform, which provides information about recalls, to automatically notify patients of relevant recalls.

Specific Comments

Major Comments

1. Because of the setup of this document, it is challenging to add comments or do any editing. Not sure what happened, but it treated every line as a single object when opened in Microsoft Word. Please check your formatting.
2. On page 2, within the abstract, under Background, there is an error in the formatting. There should be a section that begins with Aim. Instead, that section is folded into the Background section and needs to be corrected.
3. On page 8, with the MyChart message, I can see why patients felt too much wording was in this layout. Surprisingly, the Patient Advisory Council agreed to this layout and the wordiness. The focus must be on the patient’s needs, not what the FDA requires. We all have seen the Prescribers' Digital Reference, and we know that the information is too dense and too small. This is similar to that in terms of format. Enlarge the font, eliminate extraneous information, and only include information that is important to the patient and in simple English. This should be pretty feasible in the formatting of the Health Citizen and/or the MyChart message.
4. You identified problems and that patients would feel obligated to contact their provider regarding the recall. Instead of exploring how to address this so that patients wouldn’t do that, thereby increasing the significant workload on the provider’s health care team, you simply gave up. I think you could have done much more with this than say, “oh, it can’t be done.” How could you word the MyChart to direct the patients only to the pharmacy that dispenses their medication instead of the primary care provider? If you didn’t ask that question, you should have. This is not the time to give up. It’s time to inquire more to find the right answers so that this could move forward and better serve both the patients and their providers.
5. It is certainly possible, given the technical requirements to create this capability, that you ran out of time and money. However, you can still benefit your team and others by focusing on the lessons learned and how you would go forward with another study.
6. One of the things that you did not do is a first round of qualitative testing and using that feedback to make changes and do a second round. Per Nielsen [2], you only need about 5 test subjects per round to get the desired, usable results. What was preventing you from doing that? Put that in the manuscript as a limitation in your Discussion.
7. Also, on page 11, in the last paragraph of the page under Discussion, there is a comment regarding patients expecting their providers to know when a recall has occurred; I think we all know this is an unreasonable expectation. Part of the communication with the MyChart message is to inform the patients not to call their provider but to call the pharmacy that dispenses their medication, which should be right on the bottle. Again, one component of the MyChart portal messaging system, as well as any other portal messaging system, is to keep patients informed and educate them. That

should be a focus of this project, just as much as the technical components.

8. On page 12, in the last full paragraph on the page, you make a statement regarding the project that a strong case can be made for requiring each pill bottle to include the lot number (maybe) and National Drug Code of the pills. Since the FDA was a component of this project, that should probably have been something you recommended for the FDA to require and not leave to the state boards of pharmacy, as then you would get a patchwork of regulations. This would require the FDA to say that lot numbers and National Drug Codes are required on the bottles of all medications with an appropriate implementation period to allow for appropriate software and hardware adjustments. That is just as valuable a recommendation out of the study as any other.

Round 2 Review

General Comments

This paper describes a small qualitative study that aims to leverage the FDA's Healthy Citizen prototype platform, which provides information about recalls, to notify patients of relevant recalls automatically. The project team deemed the goal unattainable and provided limited lessons learned and recommendations for potential advocacy/future solutions.

Specific Comments

Major Comments

1. On page 11, in the section/paragraph beginning with "Major thematic findings included...": these are some of the lessons learned that I mentioned in my feedback.
2. On page 12, in the paragraph beginning with "The project team concluded that...": The "project team" felt this. Did the Patient Advisory Council and the test subjects share the same feeling?
3. On page 13, in the second paragraph on the page, in the sentence beginning with "Note that the FDA does not...": this would clearly be a lesson learned and could be advocated for via Congress and the Department of Health and Human Services.
4. On page 13, in the second paragraph, the next sentence, beginning with "The manufacturer and lot number of dispensed medications...": agreed. See previous comment.

Conflicts of Interest

None declared.

References

1. Gadgil M, Pavlakos R, Carini S, et al. Automating individualized notification of drug recalls to patients: complex challenges and qualitative evaluation. JMIRx Med 2026;7:e68345. [doi: 10.2196/68345]
2. Nielsen Norman Group. Why you only need to test with 5 users. URL: https://www.nngroup.com/articles/why-you-only-need-to-test-with-5-users/ [accessed 2025-10-14]

Abbreviations

FDA: Food and Drug Administration

Edited by CN Hang; submitted 18.Aug.2025; this is a non–peer-reviewed article;accepted 18.Aug.2025; published 13.Jan.2026.

Please cite as:
Marshall RC
Peer Review of "Automating Individualized Notification of Drug Recalls to Patients: Complex Challenges and Qualitative Evaluation"
JMIRx Med 2026;7:e82612
URL: https://xmed.jmir.org/2026/1/e82612
doi:10.2196/82612

Peer Review of “Use of a Specialist Telephone Consultation Line for Long COVID in Primary Care in British Columbia: Retrospective Descriptive Quality Improvement Study”

Anonymous

Related Articles:

Companion article: https://www.medrxiv.org/content/10.1101/2024.01.28.24301875

Companion article: https://med.jmirx.org/2026/1/e89710

Companion article: https://med.jmirx.org/2026/1/e57021

(JMIRx Med 2026;7:e90935) doi:10.2196/90935

KEYWORDS

internal medicine; long COVID; COVID-19; SARS-CoV-2; GP; general practice; general practitioner; consult; respiratory; infectious; respiration; primary care; telephone; telehealth

This is the peer-review report for “Use of a Specialist Telephone Consultation Line for Long COVID in Primary Care in British Columbia: Retrospective Descriptive Quality Improvement Study.”

Round 1 Review

With all due respect to the authors for their work, I personally found the language of the article [1] to be fair, but there are the following problems:

1. The study design was relatively simple, with only age and gender collected for basic characteristics and no mention of past medical history, which had a greater impact on the study results, especially since the study results showed a high rate of reported respiratory symptoms. In addition, the 40‐49 year age group also had a high prevalence of chronic respiratory illnesses; previous respiratory illnesses are bound to worsen to varying degrees after a COVID infection. Despite the high probability of missing visits or ambiguous data, the collection of past medical history is something that I personally feel should have been added, and missing data need to be accounted for.
2. As a quality improvement study, I believe that the original COVID–general internal medicine–Post-Infection Care Rapid Access to Consultative Expertise line should be introduced (such as through flowcharts) to identify problems in the follow-up process and problems affecting the results of the study and to propose more specific improvement measures such as special training for follow-up personnel to guide the enrolled patients to more accurately provide the information needed for the study.
3. There are too many confounding factors affecting the results, and the author team does not seem to have mentioned measures to minimize the impact of confounding factors on the results of the study.

Conflicts of Interest

None declared.

Reference

1. Kaushal S, Bhandal J, Birks P. Use of a Specialist Telephone Consultation Line for Long COVID in Primary Care in British Columbia: Retrospective Descriptive Quality Improvement Study. JMIRx Med 2026;7:e57021. [doi: 10.2196/57021]

Edited by A Schwartz; submitted 06.Jan.2026; this is a non–peer-reviewed article;accepted 06.Jan.2026; published 10.Feb.2026.

Please cite as:
Anonymous
Peer Review of "Use of a Specialist Telephone Consultation Line for Long COVID in Primary Care in British Columbia: Retrospective Descriptive Quality Improvement Study"
JMIRx Med 2026;7:e90935
URL: https://xmed.jmir.org/2026/1/e90935
doi:10.2196/90935

Peer Review of “Use of a Specialist Telephone Consultation Line for Long COVID in Primary Care in British Columbia: Retrospective Descriptive Quality Improvement Study”

Saidi Olayinka Olalere, MSc

Georgia Southern University, Statesboro, GA, United States

Related Articles:

Companion article: https://www.medrxiv.org/content/10.1101/2024.01.28.24301875

Companion article: https://med.jmirx.org/2026/1/e89710

Companion article: https://med.jmirx.org/2026/1/e57021

(JMIRx Med 2026;7:e89735) doi:10.2196/89735

KEYWORDS

internal medicine; long COVID; COVID-19; SARS-CoV-2; GP; general practice; general practitioner; consult; respiratory; infectious; respiration; primary care; telephone; telehealth

This is the peer-review report for “Use of a Specialist Telephone Consultation Line for Long COVID in Primary Care in British Columbia: Retrospective Descriptive Quality Improvement Study.”

Round 1 Review

General Comments

This paper [1] is structured well, with an analysis of the data obtained. The authors need to present how this data will assist the province with some figures or data.

Specific Comments

Major Comments

1. The authors mention that 6 calls were excluded but never gave an analysis of the trend of the calls.
2. Can the 6 calls drive some conclusions that can assist with the paper?
3. Can the authors give a trend line for the period of these calls and if there are related cases of different calls?

Conflicts of Interest

None declared.

Reference

1. Kaushal S, Bhandal J, Birks P, et al. Use of a Specialist Telephone Consultation Line for Long COVID in Primary Care in British Columbia: Retrospective Descriptive Quality Improvement Study. JMIRx Med 2026;7:e57021. [doi: 10.2196/57021]

Edited by A Schwartz; submitted 16.Dec.2025; this is a non–peer-reviewed article;accepted 16.Dec.2025; published 10.Feb.2026.

Please cite as:
Olalere SO
Peer Review of “Use of a Specialist Telephone Consultation Line for Long COVID in Primary Care in British Columbia: Retrospective Descriptive Quality Improvement Study”
JMIRx Med 2026;7:e89735
URL: https://xmed.jmir.org/2026/1/e89735
doi:10.2196/89735

Peer Review for “Administration Technique of Intranasal Corticosteroid Sprays Among Nepali Pharmacists: Cross-Sectional Study”

Ravi P Shankar

Manipal College of Medical Sciences, Pokhara, Nepal

Related Articles:

Companion article: http://preprints.jmir.org/preprint/83042

Companion article: https://med.jmirx.org/2026/1/e91445

Companion article: https://med.jmirx.org/2026/1/e83042

(JMIRx Med 2026;7:e91443) doi:10.2196/91443

KEYWORDS

intranasal corticosteroid spray; allergic rhinitis; device use technique; pharmacist; patient counselling, continuing pharmacy education

This is a peer-review report for “Administration Technique of Intranasal Corticosteroid Sprays Among Nepali Pharmacists: Cross-Sectional Study.”

Round 1 Review

General Comments

This is an important and well-written study [1]. My suggestions are listed below.

Specific Comments

There are some problems with language and with unnecessary capitalization of words.

Page 3: INCS sprays should be defined in full on first mention in the text.

Page 8: Can details of the ethical committee that provided the approval be provided?

Was the informed consent obtained in writing?

Scoring system: Should the crucial steps not be provided with greater marks compared to the other steps?

Page 17: Please explain the classification tree (Chi-square automatic interaction detection method) for the benefit of the readers.

Page 17: “This research is one of a kind, conducted in Nepal.” Can this sentence be modified?

Page 19: Instead of continuing medical education (CME), continuing pharmacy education (CPE) may be a better term.

Page 20: What educational aids are you referring to?

Are the educational leaflets available in the Nepali language?

Page 20: “In our study, both the increasing age (>26 y old) were significantly associated with improved INCS [intranasal corticosteroid] counseling proficiency.” This sentence mentions both but then highlights only one factor.

Was this study conducted only in Kathmandu city and not in Lalitpur or Bhaktapur?

Page 21, Limitations section: Some of the findings may be extreme due to small subgroups or model overfitting. Can this be explained?

Different fonts are used in different locations, and this should be corrected.

Conflicts of Interest

None declared.

Reference

1. Chaudhary AP, Thakur S, Sah SK. Administration technique of intranasal corticosteroid sprays among Nepali pharmacists: cross-sectional study. JMIRx Med 2026;7:e83042. [doi: 10.2196/preprints.83042]

Edited by A Schwartz; submitted 14.Jan.2026; this is a non–peer-reviewed article;accepted 14.Jan.2026; published 29.Jan.2026.

Please cite as:
Shankar RP
Peer Review for "Administration Technique of Intranasal Corticosteroid Sprays Among Nepali Pharmacists: Cross-Sectional Study"
JMIRx Med 2026;7:e91443
URL: https://xmed.jmir.org/2026/1/e91443
doi:10.2196/91443

Peer Review for “Administration Technique of Intranasal Corticosteroid Sprays Among Nepali Pharmacists: Cross-Sectional Study”

Sunny Chi Lik Au

Tung Wah Eastern Hospital, So Kon Po, China (Hong Kong)

Related Articles:

Companion article: http://preprints.jmir.org/preprint/83042

Companion article: https://med.jmirx.org/2026/1/e91445

Companion article: https://med.jmirx.org/2026/1/e83042

(JMIRx Med 2026;7:e91439) doi:10.2196/91439

KEYWORDS

intranasal corticosteroid spray; allergic rhinitis; device use technique; pharmacist; patient counselling, continuing pharmacy education

This is a peer-review report for “Administration Technique of Intranasal Corticosteroid Sprays Among Nepali Pharmacists: Cross-Sectional Study.”

Round 1 Review

General Comments

This paper [1] addresses an important gap by evaluating pharmacists’ proficiency in demonstrating intranasal corticosteroid technique, using a standardized 12-step checklist with 5 critical steps. The sample size (n=365) is reasonable for a local study, and the use of multivariate logistic regression and Chi-square automatic interaction detection decision tree analysis adds analytical depth. The findings highlight systemic issues, such as inadequate training and curriculum gaps, which could inform policy changes to improve allergic rhinitis management and reduce adverse effects like epistaxis.

Specific Comments

Major Comments

1. Simple random sampling was used for pharmacies, but details on how wards were selected or how pharmacists within pharmacies were approached are vague. Please supplement and elaborate on further details of the randomization. More information on such would help lower the selection bias (eg, busier or more accessible pharmacies might be overrepresented).
2. The questionnaire’s validity is only face-validated by experts, with no content or construct validity testing mentioned. Reliability was assessed via Cronbach alpha (0.758) on a small pilot (n=15), which is acceptable but not robust. The cutoff for “adequate” proficiency (>6/12 marks) is based on the median score and expert opinion, which feels arbitrary and not clinically validated. Why not base it on critical steps alone, given their emphasis on efficacy and safety? Only 6% performed all 5 critical steps correctly, yet 47% were deemed “adequate” overall. This discrepancy suggests the threshold may be too lenient, masking true incompetence in high-impact areas like directing the nozzle away from the septum (to prevent epistaxis) or exhaling through the mouth (to optimize deposition). Please address these in the Discussion section.
3. Self-reported variables (eg, counseling frequency, use of materials) are prone to recall or social desirability bias, especially in an in-person interview setting. Please supplement these in the Discussion section.
4. The multivariate binary logistic regression identifies associations (eg, male gender, older age, higher qualifications linked to better proficiency), but potential confounders like pharmacy type (independent vs chain) or workload details are not controlled for. Odds ratios are extreme in places (eg, BPharm holders 97% less likely to perform inadequately, or frequent counselors 11 times more proficient), which may stem from small subgroups or multicollinearity.
5. Gender differences (males ~2 times more proficient) were found but underlying factors were not explored (eg, access to workshops, cultural biases). Please elaborate more or address the potential underlying factors in the Discussion section.
6. “Educational materials” are linked to better proficiency, but what constitutes these (eg, leaflets, videos)? Please specify for readers to enhance the proficiency on applying the study’s results.
7. Reference 16 has the wrong format for the volume, issue, and page numbers:Al-Taie A. A Systematic Review for Improper Application of Nasal Spray in Allergic Rhinitis: A Proposed Role of Community Pharmacist for Patient

Education and Counseling in Practical Setting. Asia Pacific Allergy. 2025:10-5415.The full information from PubMed is as below:Al-Taie A. A systematic review for improper application of nasal spray in allergic rhinitis: A proposed role of community pharmacist for patient education and counseling in practical setting. Asia Pac Allergy. 2025 Mar;15(1):29-35. doi: 10.5415/apallergy.0000000000000173. Epub 2025 Jan 13. PMID: 40051424; PMCID: PMC11882221.Therefore, "2025:10-5415" should be "2025 Mar;15(1):29-35."Please revise the whole reference list to see if any other typos exist.

Conflicts of Interest

None declared.

Reference

1. Chaudhary AP, Thakur S, Sah SK. Administration technique of intranasal corticosteroid sprays among Nepali pharmacists: cross-sectional study. JMIRx Med 2026. [doi: 10.2196/preprints.83042]

Edited by A Schwartz; submitted 14.Jan.2026; this is a non–peer-reviewed article;accepted 14.Jan.2026; published 29.Jan.2026.

Please cite as:
Au SCL
Peer Review for "Administration Technique of Intranasal Corticosteroid Sprays Among Nepali Pharmacists: Cross-Sectional Study"
JMIRx Med 2026;7:e91439
URL: https://xmed.jmir.org/2026/1/e91439
doi:10.2196/91439

Author's Response to Peer Review Reports on "Investigating the Variable Component of the Systematic Error, a Neglected Error Parameter: Theoretical Reevaluation Study"

Atilla Barna Vandra, MS

Spitalul Clinic Judetean de Urgenta Brasov, Str. Berzei 2 Bl. B. ap 20, Brasov, Romania

Corresponding Author:
Atilla Barna Vandra, MS
Spitalul Clinic Judetean de Urgenta Brasov, Str. Berzei 2 Bl. B. ap 20, Brasov, Romania

Related Articles:

Companion article: https://www.medrxiv.org/content/10.1101/2023.05.24.23290382v1

Companion article: https://med.jmirx.org/2026/1/e88830

Companion article: https://med.jmirx.org/2026/1/e90221

Companion article: https://med.jmirx.org/2026/1/e49657

(JMIRx Med 2026;7:e88981) doi:10.2196/88981

KEYWORDS

repeatability condition; reproducibility within laboratory condition, measurement; systematic error; clinical laboratory; quality control; bias; QC; statistical; statistics; mathematics; computer simulation; standard deviation

This is the author's response to peer-review reports on "Investigating the Variable Component of the Systematic Error, a Neglected Error Parameter: Theoretical Reevaluation Study."

Round 1 Review

Reviewer C [1]

1. The handling of constant or intermittent bias has been a challenge for more than 200 years, especially since Gauss and Laplace's work in the early 19th century. The author of this paper [2] refers to Eisenhart's excellent 1963 paper, which is appropriate but not as the origin of the variable component of the systematic error (VCSE). Shewhart's 1923 and 1939 books [3,4] also address this matter.

Response: Thanks for the information about the first constant or variable bias debates. I have changed the text referring to C Eisenhart as the first author, who mentioned VCSE, referring to Shewhart, mentioning him in the reference list.

2. Bias, including VCSE, is also a major point of contention in the International Bureau of Weights and Measures/International Organization for Standardization work on the Guide to the Expression of Uncertainty in Measurement (GUM) and its revision. The complexity of the issues is illustrated, for example, in a book by Krystek [5]. The current manuscript illustrates the opinions of its author but fails to illustrate the background of the immense scientific literature and debates that have already dealt with the matter.

Response: I have found in JCGM (Joint Committee for Guides in Metrology) GUM-6:2020 paragraph 10.6 a discussion about drift effects and compared it with JCGM 100:2008 GUM 3.2.4 recommendations, which hide the assumption of a constant bias. Neither a correction (B.2.23) nor a correction factor (B.2.24) can eliminate a function (a time-variable bias). Please take a look at Principle 3 in the new version. I have downloaded and read the study by Krystek and refer to it in the new version.

3. The author needs to clarify whether he adheres to the error or uncertainty paradigms in measurement uncertainty/error questions. The current manuscript represents a mixture of both.

Response: I discuss in the new version that the total measurement error (TE) and uncertainty of measurement (MU) paradigms are, in essence, complementary, not contradictory. (This was also mentioned in the old version; I tried to be more explicit). TE and MU are linked to two different points of view. The TE approach cannot challenge the MU in calculating the MU, nor can the MU approach challenge TE in short-term decisions, as in internal quality control decisions. More than that, the two approaches have common parts, being based on the same oversimplified error model, and corrections influence both theories.

4. The "thought-provoking and even shocking" fact that the Westgard rules and power calculations are based on repeatability uncertainty and not on reproducibility uncertainty is well-known by metrologists in clinical chemistry. Unfortunately, mentioning this fact commonly hurts the

sentiments of a majority in our field, and many of us avoid harping on it. A prerequisite for appropriately using Westgard rules and variants is that changing goal mean values are used as calibrators, and reagent lots change over time.

Response: Quoting his text: "The 'thought-provoking and even shocking' fact that the Westgard rules and power calculations are based on repeatability uncertainty and not on reproducibility uncertainty is well-known by metrologists in clinical chemistry." I have reformulated, eliminating the label, referring only to the recommendations of Westgard et al [6] to design the Levey-Jennings graphs based on the SD calculated from long-term control data ("at least 20 measurements over at least two weeks or ten working days, and preferably over at least four weeks or 20 working days"). I did mention the CLSI C24-Ed4 2016 recommendations: "From a clinical point of view, repeatability is rarely of interest. Generally, within-laboratory precision estimates are clinically more relevant because they reflect variability over time intervals somewhat more representative of intervals between repeat measurements for a patient being monitored for chronic disease." Also: "The Sigma metric may be calculated using either repeatability or within-laboratory imprecision as the estimate of the SD. However, for the most useful estimates of the Sigma metric, the within-laboratory SD is the best choice." I also mentioned the paper of Westgard and Groth [7], in which they accept that the power function graphs are designed with s_r (SD measured in constant, repeatability conditions). Seemingly, they knew the contradiction between the Levey-Jennings graphs and the power function graphs but continued to suggest creating the Levey-Jennings graphs with sRW (SD measured in variable, reproducibility within laboratory conditions).

5. The traceability hierarchies used when producing reference materials and calibrators are usually claimed to explain the variations experiences (eg, during lot-number changes). The author apparently does not accept this explananation of the main cause of lot-number shifts/bias, and he needs to explain why his mathematical/statistical theory should be accepted instead.

Response: I did not state that I do not accept that the nominal value errors (with the words of reviewer C: "The traceability hierarchies used when producing reference materials and calibrators" are not sources of "the variations experiences [eg, during lot-number changes]") (with other words of the VCSE(t) [variable component of systematic error at the moment t]). More than that, I have sustained that shifts caused by the calibrations are one of the causes of the VCSE(t). Lot number changes are only one of the causes of shifts after calibrations. The revised manuscript also details other causes, such as the measurement error during calibration and the reconstitution errors linked to the reference material bottles. My theory does not contradict the importance of the nominal value errors; more than that, it completes it. I have stated: In the time frames between human interventions, the bias variations are predictable, hidden behind the noise of the RE (random error component). A lot of change is a human intervention. This is one of the causes of the bias variation. Both lot number changes and control bottle changes (reconstitution error) cause shifts in bias.

6. In a crucial part of his manuscript, the author claims that "While RE changes unpredictably from measurement to measurement, VCSE(t) remains quasi-constant in a given day, influencing all measurement results obtained in that day systematically. But in long-term experiments, VCSE(t) becomes a cyclical time-variable function, which repeats the same values after unequal periods. (A period may last even one month)." The author presents Cobas 6000 analyzer data in support of his thesis. However, data from a variety of measuring systems, lot changes, and measurands are needed before this theory of a cyclical phenomenon is chosen instead of a theory of random components.

Response: I have detailed the idea of cyclical variation in the revised version. I hope I did it more explicitly. The reagent property changes are unidirectional, causing drifts. The bias increase (in absolute values) cannot continue endlessly because shifts caused by human interventions correct it (reagent changes, calibrations). The consequence is a sawtooth-like graph, a cyclical variation as in Figure 3. I have substituted Figure 3 (preprint version) with two others (one real-life), explaining that only in the case of significant drifts can the sawtooth-like character be observed behind the noise of the random errors. Without a known cause, the causes of SD increase in time were labeled "random." However, we cannot identify any causes of variable RE (the sRW variations are caused not by RE but by the VCSE(t)). I have underlined that the presented phenomenology was observed on all analyzers I have worked with. Also, I highlighted that the given examples can be visually observed only if they are significant (usually, the phenomena are hidden behind the RE). Because the real cause was unknown, the myth of random bias variations was born.

Anonymous [8]

1. The study does not provide empirical confirmation of suggested approaches using real-world data, although it mentions computer simulations and experimental verification. The suggested methodologies' efficacy and dependability are yet unknown in the absence of empirical validation.

Response: Thanks for the idea. I did not want to introduce more graphs because of the length of the study, but in the revised version, I shall do it.

2. Linear drifts in daily means across time are assumed in the study. Numerous factors, such as the environment, instrument calibration, and reagent stability, can affect real-world drift patterns and lead to nonlinear trends in daily means over time. The study might have simplified the complicated nature of drift processes by assuming linearity, which could result in estimates of mean values and error components that are not true.

Response: Thanks for the idea. In the revised version, I shall explain why environmental changes have an insignificant influence on thermostated reactions in an automatized laboratory with air conditioning. "Human (operator) errors" and "laboratory errors" are redundant in the error list because they always act via instrumental and noninstrumental errors. Bias variations are always specific to the reaction; therefore, they may have only noninstrumental causes (reagent stability and calibration curve changes). The quasi-linear drifts are caused only by the reagent

property changes. Random variations in reagent properties would contradict the laws of chemistry. Calibrations cause shifts, which have known moments (human interventions) but to a random extent. The "linear drifts in daily means" happen in the intervals of the human interventions (calibrations, regent changes, control bottle changes). Across these, we cannot apply the quality control rules. This is also the recommendation of Westgard. However, there are unexpected shifts (eg, caused by carry-over phenomena), and the quality control must be able to detect them.

3. The assumption that information from internal quality control sources alone can be used to accurately calculate VCSE(t) is inaccurate. Even though internal quality control data offer insightful information on short-term variability, they might not include all sources of variation, particularly those pertaining to outside variables like environmental shifts, instrument performance, or operator technique. Ignoring these outside influences could result in an inaccurate or understated VCSE(t), which would compromise the validity of the suggested quality control techniques.

Response: Thanks for the observation. In the revised version, I shall provide details (see the former answer to observation 2). The anonymous reviewer is right; VCSE(t), by definition, is variable, but so is sVCSE. VCSE(t), the value of the VCSE in the moment t, can be determined accurately within the limits of the statistical methods, but it only has 24-hour validity. I agree that its determination has a high cost/effectiveness ratio. Its approximate evaluation in the internal quality control is a better choice. sVCSE depends on the time frame. Therefore, its determination has acceptable accuracy only from yearly data. The method was described. Because sVCSE depends on the time frame, its value cannot be used only in the same time frame in which it was determined.

As will be detailed more in the revised version, the importance of the VCSE(t) and sVCSE is not their value but the knowledge about their existence. We can avoid redundant use by highlighting these error components in equations. In calculations, they are summed with other error components (ie, are "hidden" in Br(t) or sRW); therefore, their absolute values are not important. However, it is essential to avoid the use of Br(t) and sRW in the same equations (eg, TE = Br(t) + z*sRW).

4. Although there is a suggestion in the Conclusions section that the present quality control paradigm needs to be revised, there is no concrete plan or set of recommendations based on statistical or mathematical concepts.

Response: Thanks. I will make recommendations based on the concepts presented in the study to improve the quality control. The next step is to present a new quality control system.

5. The paper lacks a Discussion section, which could have allowed the author to interpret and contextualize the study's findings. Additionally, it could have provided an opportunity to compare the study's findings with previous studies, discuss their implications, and address potential sources of error or bias.

Response: The revised version will have a more detailed Discussion section, and the paper will be reorganized. The implications were discussed: separating the bias components prevents the VCSE(t) redundancy in equations and suggested corrected equations. The revised version will be more explicit in presenting the bias sources: the reagent instability and the calibration graph errors. I also included a Comparison with the Literature section.

Round 2 Review

Reviewer C

In the first round of reviews, I asked for "(1) a well-structured manuscript based on (2) extensive knowledge of the state of the art in calculating measurement uncertainty and (3) well-written English text."

The revised version of the manuscript has improved the English text but needs to improve in the two other aspects.

I agree that the paper's subject is essential. The author is well-versed in mathematical statistics and has practical experience in laboratory quality control. However, the manuscript lacks in:

1. *Counting in metrological aspects*
2. *Using a conventional manuscript structure*
3. *Showing sufficient real laboratory results and the consequence of using the proposed paradigm on real laboratory results*

Response: I shall begin with point 2.

I am quoting Reviewer C: "I asked for "(1) a well-structured manuscript…" "The manuscript lacks a conventional manuscript structure."

I have rewritten the whole manuscript, trying to respect the required structure. I thought that both the expressions and the content are important. I have given slightly different names, with the same sense. The titles of the main sections were Introduction (introduction), Methods (materials and methods), Results (experimental data and computer simulation), Discussion (discussion), Conclusions (conclusions).

Each section has subsections. I divided them into subsections to make it easier to trace the experimental data.

In the Methods section, three experiments were described, showing three different phenomena. In the Results section, each has a different subsection. The Discussion section was divided into five subsections, each discussing various aspects.

I have renamed the main sections, and I numbered the subsections. I hope it is OK now.

I am quoting Reviewer C: (I asked for) "extensive knowledge of the state of the art in calculating measurement uncertainty."

The anonymous reviewer, in the first review round, asked me to adhere to one of the paradigms: TE or UM. I answered him/her that I did not want to because the two paradigms have different areas of use and are linked to two different points of view: short- and long-term. Neither can substitute the other in its area of applicability. However, this study focuses on internal quality control decisions, which are short-term decisions, and

therefore, the study focuses on TE. However, changing the error model has consequences for both paradigms because both are built on the same error model. This theoretical study focuses on equations and mathematics, not applicability. Briefly: bias is variable; it is a time-variable function, and we must make a distinction between bias types. Otherwise, there is the risk of redundant use. Bias variability has two sources; both are noninstrumental: the reagent instability and the calibration errors. s_r is the estimator of the true RE, and sRW is erroneously considered the measure of the RE because it is the measure of all variable components (RE + VCSE). There is nothing about the uncertainty of measurement. I have mentioned several times that further studies are necessary to analyze the applicability of the new error model. These analyses neither fit the study's task nor its word count limits. The aim was to be the foundation stone for a new quality control system. I have mentioned the guiding principles, but this study is not about presenting and proving the applicability of the new system. I cannot publish a new quality control system without the new error model. I have been waiting two years to publish it, but I cannot until the error model is published. To detail the hidden false assumptions of the Westgard-rules–based quality control system, I need the help of the new error model. To detail all principles, too. I need space to present a new rule system based on a modified Levey-Jennings graph and sustain it with computer simulation data and probability calculations (Westgard et al also did not present their own quality control system with real-life data; it was only based on computer simulations). However, I have proofs based on real-life data. These cannot be presented before the former analysis and descriptions are presented. This study is a foundation stone, not a new method, and neither of these are based on UM.

I suppose, but I am not sure, that Reviewer C wants an analysis of the error model based on GUM. The UM starts with correcting the discovered biases in the first step using corrections and correction factors. But a constant cannot fix a variable!! A single bias value does not give information about the constant or proportional character of the bias. A smaller bias than its uncertainty cannot be corrected. Bias is variable because of its properties, and this variability is significant in the clinical laboratory. The external quality assessment (EQA) results are obtained after a considerable delay while calibrations and reagent changes are done. A correction applied after such a delay is risky! The UM strategy may be efficient in lifeless domains but not in the conditions of the variability of biases in the clinical laboratory. Because of the delay, we do not correct the measured bias based on the last EQA, but a significantly different one! Such corrections assume a constant bias, which is a false assumption. The analysis of such contradictions neither fits the study's aims nor its limits, mainly because Reviewer C asks for proof of real-life data. Therefore, a presentation of the UM literature, in my opinion, is unnecessary.

The error model also has consequences for UM equations. To prove that the UM equations are based on "the bias" definitional uncertainty is not about the error model, but rather, it is a review of the UM. It is not the task of this study, which is about an error model. It is not the same as correcting a bias value or a mean bias. The uncertainty of a value and a mean is different, too. The UM equations include neither the uncertainty caused by the reconstitution error of the reference material, which is usually bigger than the uncertainty of its nominal value, nor the uncertainty of the sRW (which may be double in the next month). This is not a review paper. The literature about conditions not respected for a correct EQA does not fit in a single room. The most used methods (because of the unavailability of commutable certified materials and economic reasons) are using peer group means as surrogate reference values. The formula for calculating the nominal value uncertainty and the target value is invalid in these groups. Therefore, there are two different uCref values, one declared (erroneously calculated) and the other real. The latter is included in RMSbias. RMSbias + uCref redundantly uses the uCref term (one is bottom-up, the other is an up-down parameter, which cannot be mixed without redundancy…let me continue?)

I have included only a brief (two pages) presentation of the upper analysis. However, it is a cuckoo's egg. Neither fits in the task of a theoretical study about an error model nor real-life data proofs that the corrections based on the last EQA increase the variability of biases (instead of reducing them), contributing to the increase of total uncertainty. (I have made some simulations based on real-life data (based on four-year EQA data), and this was the conclusion: only mean biases can be used in corrections, single EQA data cannot). The presentation of these does not fit in the limits of this study.

My question remains: what do I need to prove with real-life data? That bias is variable? Even if it is accepted that it is variable, I need to prove with real-life data that we must make a difference between different biases. That the variable bias does not fit into the classical error model? That calibrations and reagent property changes cause bias variations? All these were described in the literature in mosaic pieces. A quality control system based on s_r and the avoidance of alarms in the case of incorrigible biases USING DIFFERENT RULES, not the actual Westgard rules, and the fact that correctly applied Westgard rules are an unfunctional quality control system, does not fit into the limits of this study.

I am quoting Prof A Marusteri: "If this is true, nothing we thought certain points in quality control are unquestionable anymore." With such a radical change, several phenomena change in quality control. Each needs proof. However, so many things cannot be presented in a study. Therefore, I have mentioned several times that something "needs a separate study."

3. "However, the manuscript lacks in "Counting in metrological aspects."

Response: I do not understand this critique; I do not understand what the statement refers to.

References

1. Theodorsson E. Peer review of “Investigating the Variable Component of the Systematic Error, a Neglected Error Parameter: Theoretical Reevaluation Study”. JMIRx Med 2026;7:e88830. [doi: 10.2196/88830]
2. Vandra AB. Investigating the variable component of the systematic error, a neglected error parameter: theoretical reevaluation study. JMIRx Med 2026;7:e49657. [doi: 10.2196/49657]
3. Shewhart WA. Economic Control of Quality of Manufactured Product: D. Van Nostrand Company; 1923.
4. Shewhart WA. Statistical Method from the Viewpoint of Quality Control: Dover Publications; 1939.
5. Krystek M. Calculating Measurement Uncertainties: Beuth Verlag GmbH; 2016.
6. Westgard JO, Barry PL, Hunt MR, Groth T. A multi-rule Shewhart chart for quality control in clinical chemistry. Clin Chem 1981 Mar;27(3):493-501. [Medline: 7471403]
7. Westgard JO, Groth T. Power functions for statistical control rules. Clin Chem 1979 Jun;25(6):863-869. [Medline: 445821]
8. Anonymous. Peer review of “Investigating the Variable Component of the Systematic Error, a Neglected Error Parameter: Theoretical Reevaluation Study”. JMIRx Med 2026;7:e90221. [doi: 10.2196/90221]

Abbreviations

EQA: external quality assessment
GUM: Guide to the Expression of Uncertainty in Measurement
JCGM: Joint Committee for Guides in Metrology
MU: uncertainty of measurement
RE: random error component
s_r: SD measured in constant, repeatability conditions
sRW: SD measured in variable, reproducibility within laboratory conditions
TE: total measurement error
VCSE: variable component of the systematic error
VCSE(t): variable component of systematic error at the moment t

Edited by T Leung; submitted 04.Dec.2025; this is a non–peer-reviewed article;accepted 04.Dec.2025; published 27.Feb.2026.

Please cite as:
Vandra AB
Author’s Response to Peer Review Reports on “Investigating the Variable Component of the Systematic Error, a Neglected Error Parameter: Theoretical Reevaluation Study”
JMIRx Med 2026;7:e88981
URL: https://xmed.jmir.org/2026/1/e88981
doi:10.2196/88981

Authors' Response to Peer Review of "Associations Between IT Job Stressors and Anxiety, Depression, and Stress: Cross-Sectional Study"

Edlin Garcia Colato[1], MPH, PhD; Nianjun Liu[2], PhD; Angela Chow[3], PhD; Catherine M Sherwood-Laughlin[3], HSD, MPH; Jonathan T Macy[3], MPH, PhD

[1]Department of Health and Wellness Design, School of Public Health, Indiana University Bloomington, 1025 E 7th St, Bloomington, IN, United States
[2]Department of Epidemiology and Biostatistics, School of Public Health, Indiana University Bloomington, Bloomington, IN, United States
[3]Department of Applied Health Science, School of Public Health, Indiana University Bloomington, Bloomington, IN, United States

Corresponding Author:
Edlin Garcia Colato, MPH, PhD
Department of Health and Wellness Design, School of Public Health, Indiana University Bloomington, 1025 E 7th St, Bloomington, IN, United States

Related Articles:

Companion article: https://www.medrxiv.org/content/10.1101/2025.02.24.25322785v1

Companion article: https://med.jmirx.org/2026/1/e91383

Companion article: https://med.jmirx.org/2026/1/e73211

(JMIRx Med 2026;7:e91437) doi:10.2196/91437

KEYWORDS

survey; association; occupational health; mental health; stressors; IT; IT professionals; United States; workplace; depression; anxiety; stress; help-seeking; health literacy

This is the authors' response to the peer review of "Associations Between IT Job Stressors and Anxiety, Depression, and Stress: Cross-Sectional Study." [1]

Round 1 Review

Reviewer AD [2]

General Comments

Although the data basis of this study is not very reliable due to some methodological limitations (cross-sectional design, online survey, self-reporting only, self-selection bias), the study does provide some interesting insights. I have the following comments to make.

Specific Comments

Major Comments

What I miss most is a more differentiated discussion of the assumed mediation by mental health literacy (MHL). In my view, it is not necessarily plausible. It would be just as plausible to assume that people with high MHL are more competent in knowing what to do by themselves. In addition, MHL itself can also have a protective function, in that people with high MHL also know what they can do by themselves to protect themselves in the event of a high workload and apply any compensatory measures (eg, relaxation).

Response: Thank you for your thoughtful feedback regarding the assumed mediation by MHL. We appreciate your point that the plausibility of MHL as a mediator warrants a more nuanced discussion. In the revised manuscript, we have expanded this section to acknowledge the limitations in interpretation.

Minor comments

1. Some information on the analysis should be added to the abstract.

Response: We have updated it to include the following: "Descriptive statistics, regression models, and mediation analyses were conducted for CESD-10, GAD-7, and PSS-10."

2. "Anxiety" should also be included in the keywords.

Response: Thank you for the suggestion, we have included "anxiety" as a keyword.

3. With regard to the excluded cases, no cases are mentioned that were excluded due to conspicuous response behavior (eg, monotonous patterns) or too rapid completion ("speeders"). Why?

Response: We are aware that excluding cases for monotonous patterns or too-rapid completion is a common practice to ensure data validity. We have updated our text to explain that rapidly completed surveys were marked as invalid and were thus removed from the final sample. However, in our review of our

data, there was no evidence of monotonous completion. All retained observations had completion times that were within the expected ranges, and the response patterns showed sufficient variability.

We added the following text: "Review of the data showed no evidence of conspicuous response behavior. Outliers in average completion time of the survey that showed the survey was completed in only a few minutes however were excluded. A total of 388 (84.3%) of the remaining 460, who provided consent and were determined to be valid responses, completed the survey."

4. Please add information on how many people in total were initially contacted.

Response: We have added that approximately 2336 individuals were contacted.

5. A core element of the study is the initial identification of IT-specific stressors. Here, further information on the criteria for the selection of the experts and the methodological procedure for capturing the stressors is essential (interviews? focus group? workshop?).

Response: We added the following text: "IT experts were selected based on predefined criteria, including their professional qualifications and practical experience. Interviews were conducted with the IT experts, allowing for the development of a comprehensive list of stressors."

6. The study just assessed the intention to seek help. Did it also assess whether professional help had been sought in the past 12 months? If not, why not? This would have been very easy to capture and a much more reliable criterion than just intentions.

Response: Our study focused on assessing intention to seek professional help as a proxy for help-seeking behavior, which aligns with the theoretical framework underpinning our research. We did not include a measure of actual help-seeking within the past 12 months. The main reason was that the objective was to examine prospective behavioral tendencies rather than retrospective behaviors. While we acknowledge that past behavior is an important factor, our design prioritized capturing motivational aspects to inform future help-seeking decisions.

7. Please add a table with the most important information about the sample.

Response: We added back the sociodemographic table, Table 1 ("Characteristics by Sex"), in the Results section.

8. Please also add a table with the frequencies of the individual stressors as well as the distribution of multiple stressors (ie, how often people reported 1 stressor, 2 stressors, etc, up to 12 stressors). This is also interesting, as the effect of multiple stressors appears to be surprisingly small. The type of stressor therefore seems to be more decisive than the frequency/diversity.

Response: Thank you for this suggestion, we have created the requested table and have included it in the appendices as supplemental material.

9. Table 1 has a different font type, please adjust.

Response: Thanks for pointing this out. We removed the original Table 1 and instead provided the odds ratios, CIs, and *P* values in the text, and we have reviewed the other tables to ensure they use consistent font types.

10. Some of the terms in Table 2 are written inconsistently (eg, "p-value"/"P value").

Response: We have updated all mentions of *P* values throughout the tables and text to ensure consistency.

11. Please add a legend beneath Table 2, explaining the abbreviations of the measures.

Response: Thank you, we have updated the table to include full definitions of abbreviations of the measures, as suggested.

12. The discussion basically only addresses why mediation shows no effect with regard to stress, but not why this is also the case with anxiety. Please also address this.

Response: Thank you for bringing this to our attention. We revised the discussion to include possible reasons why mediation was not observed for anxiety. We added the following text: "The lack of mediation found between MHL and anxiety and stress could be due to the measurement timing or the cross-sectional design that limited the ability to detect indirect effects for both anxiety and stress. Future research using longitudinal designs and alternative mediators could clarify whether these null findings reflect a true absence of mediation or methodological constraints."

13. Incidentally, the mediation effect for depression should not be overestimated. The effect just tips significance and the size of the indirect effect is rather small relative to the huge direct effect.

Response: Thank you for highlighting this point. We agree that the mediation effect for depression should not be overestimated. The mediation suggests a pathway worth noting. We updated the text to clarify that the indirect effect is modest, as follows: "The mediation results suggests a pathway worth noting between MHL and help-seeking for depression; although the indirect effect is modest. MHL had only a partial mediation for depression, but not for anxiety or stress."

14. The assessment of MHL by self-report is not necessarily a limitation per se, as there are both objective and subjective concepts in MHL. The question is rather which version is the more suitable for operationalization for testing your hypotheses.

Response: Thank you for pointing that out. MHL can be conceptualized both objectively and subjectively. We chose a self-report measure because our hypotheses focused on individuals' perceived ability to recognize and respond to mental health issues, which aligns with the subjective dimension of MHL. However, we agree that discussing why this operationalization was most suitable would strengthen the manuscript. We specifically stated that "MHL-W is the only validated MHL instrument that is specific to the workplace making it the most suitable for this type of research."

15. The references are still very inconsistent, eg, some journal titles are abbreviated/others are not; some references lack information (eg, Northwave—Where did "In" appear?) or the

year of publication (eg, for Boehm et al [3]). Please check the reference list manually throughout.

Response: All references have been updated with the *JMIRx Med* Endnote tool and manually reviewed for consistency and accurate information.

References

1. Garcia Colato E, Liu N, Chow A, Sherwood-Laughlin CM, Macy JT. Associations between IT job stressors and anxiety, depression, and stress: cross-sectional study. JMIRx Med 2026;7:e73211. [doi: 10.2196/73211]
2. Mühlan H. Peer review of "Associations Between IT Job Stressors and Anxiety, Depression, and Stress: Cross-Sectional Study. JMIRx Med 2026;7:e91383. [doi: 10.2196/91383]
3. Boehm MA, Lei QM, Lloyd RM, Prichard JR. Depression, anxiety, and tobacco use: overlapping impediments to sleep in a national sample of college students. J Am Coll Health 2016 Oct;64(7):565-574. [doi: 10.1080/07448481.2016.1205073] [Medline: 27347758]

Abbreviations

MHL: mental health literacy

Edited by A Grover; submitted 14.Jan.2026; this is a non–peer-reviewed article;accepted 14.Jan.2026; published 03.Mar.2026.

Please cite as:
Garcia Colato E, Liu N, Chow A, Sherwood-Laughlin CM, Macy JT
Authors' Response to Peer Review of "Associations Between IT Job Stressors and Anxiety, Depression, and Stress: Cross-Sectional Study"
JMIRx Med 2026;7:e91437
URL: https://xmed.jmir.org/2026/1/e91437
doi:10.2196/91437

Authors' Response to Peer Reviews of "Use of a Specialist Telephone Consultation Line for Long COVID in Primary Care in British Columbia: Retrospective Descriptive Quality Improvement Study"

Saniya Kaushal[1,2], BCh, BAO (Hons), MB; Jastinder Bhandal[3], BKin; Peter Birks[4], BMSc, MD, MHA; Jesse Greiner[4], BSc, MSc, MD, MBA; Adeera Levin[4], MD; Michelle Malbeuf[1], BSCN, MHA; Zachary Schwartz[4], BSc, MD

[1]Providence Health Care Research Institute and Provincial Health Services Authority, 1081 Burrard Street, Vancouver, BC, Canada

[2]Postgraduate Medical Education – Internal Medicine, Toronto Metropolitan University, Toronto, ON, Canada

[3]School of Medicine, University of Limerick, Limerick, Ireland

[4]Faculty of Medicine, The University of British Columbia, Vancouver, BC, Canada

Corresponding Author:
Saniya Kaushal, BCh, BAO (Hons), MB
Providence Health Care Research Institute and Provincial Health Services Authority, 1081 Burrard Street, Vancouver, BC, Canada

Related Articles:

Companion article: https://www.medrxiv.org/content/10.1101/2024.01.28.24301875

Companion article: https://med.jmirx.org/2026/1/e89735

Companion article: https://med.jmirx.org/2026/1/e90935

Companion article: https://med.jmirx.org/2026/1/e57021

(JMIRx Med 2026;7:e89710) doi:10.2196/89710

KEYWORDS

internal medicine; long COVID; COVID-19; SARS-CoV-2; GP; general practice; general practitioner; consult; respiratory; infectious; respiration; primary care; telephone; telehealth

This is the authors' response to peer-review reports for "Use of a Specialist Telephone Consultation Line for Long COVID in Primary Care in British Columbia: Retrospective Descriptive Quality Improvement Study."

Round 1 Review

Reviewer AG [1]

Major Comments

1. The authors of this study [2] *mention that 6 calls were excluded but never gave an analysis of the trend of the calls.*

Response: We now specify the reasons for excluding the 6 calls, which included unclear documentation, no discernible COVID-related question, or insufficient information. These exclusions are described in the Data Source and Call Selection subsection of the Methods. For transparency, we also clarify that excluded calls were logged to ensure consistent application of inclusion criteria.

2. Can the 6 calls drive some conclusions that can assist with the paper?

Response: As the 6 excluded calls lacked sufficient information to categorize meaningfully, their content was not analyzed to avoid introducing misclassification bias. This has been noted as a limitation in the Discussion, where we suggest that future audits could include minimal call documentation to allow sensitivity analyses.

3. Can the author give a trend line for the period of these calls and indicate if there are related cases among different calls?

Response: Temporal trends in call volume and related case patterns across pandemic phases and relative to vaccine rollout are now presented in the Results. Figure 1 illustrates these changes over time.

Anonymous [3]

Major Comments

1. The study design was relatively simple, with only age and gender collected for basic characteristics and no mention of past medical history, which had a greater impact on the study results, especially since the study results showed a high rate of reported respiratory symptoms. In addition, the 40 - 49 year age group also had a high prevalence of chronic respiratory

illnesses; previous respiratory illnesses are bound to worsen to varying degrees after a COVID infection. Despite the high probability of missing visits or ambiguous data, the collection of past medical history is something that I personally feel should have been added, and missing data need to be accounted for.

Response: We agree that the lack of past medical history is a limitation inherent to the service-level documentation available for this quality improvement project. We have now added this explicitly to the Limitations section and suggested that future Rapid Access to Consultative Expertise (RACE) audits include optional fields for past medical history and data completeness tracking to improve interpretability (Study Strengths and Limitations section of the Discussion).

2. As a quality improvement study, I believe that the original COVID–general internal medicine–Post-Infection Care RACE line should be introduced (such as through flowcharts) to identify problems in the follow-up process and problems affecting the results of the study and to propose more specific improvement measures such as special training for follow-up personnel to guide the enrolled patients to more accurately provide the information needed for the study.

Response: The Methods section now includes a concise description of the original RACE line consultation process, outlining call initiation, triage by specialists, and documentation back to primary care providers.

We also clarify how the service supports primary care provider decision-making and identify potential improvement measures, including educational resources and standardized clinical algorithms, to guide future quality improvement initiatives (Methods and Discussion sections).

3. There are too many confounding factors affecting the results, and the author team does not seem to have mentioned measures to minimize the impact of confounding factors on the results of the study.

Response: We expanded the Limitations to address key confounding factors, such as regional variation in access or awareness of the RACE line, heterogeneity in documentation quality, and evolving case definitions of long COVID. We also note that future evaluations could incorporate structured data fields and prospective data collection to reduce confounding and enhance data reliability (Study Strengths and Limitations section in the Discussion).

References

1. Olalere SO. Peer review of "Use of a Specialist Telephone Consultation Line for Long COVID in Primary Care in British Columbia: Retrospective Descriptive Quality Improvement Study". JMIRx Med 2025;7:e89735. [doi: 10.2196/89735]
2. Kaushal S, Bhandal J, Birks P, et al. Use of a Specialist Telephone Consultation Line for Long COVID in Primary Care in British Columbia: Retrospective Descriptive Quality Improvement Study. JMIRx Med 2026;7:e57021. [doi: 10.2196/57021]
3. Anonymous. Peer review of "Use of a Specialist Telephone Consultation Line for Long COVID in Primary Care in British Columbia: Retrospective Descriptive Quality Improvement Study". JMIRx Med 2026;7:e90935. [doi: 10.2196/90935]

Abbreviations

RACE: Rapid Access to Consultative Expertise

Edited by A Schwartz; submitted 16.Dec.2025; this is a non–peer-reviewed article;accepted 16.Dec.2025; published 10.Feb.2026.

Please cite as:
Kaushal S, Bhandal J, Birks P, Greiner J, Levin A, Malbeuf M, Schwartz Z
Authors' Response to Peer Reviews of "Use of a Specialist Telephone Consultation Line for Long COVID in Primary Care in British Columbia: Retrospective Descriptive Quality Improvement Study"
JMIRx Med 2026;7:e89710
URL: https://xmed.jmir.org/2026/1/e89710
doi:10.2196/89710

Authors' Response to Peer Reviews of "Awareness, Experiences, and Attitudes Toward Preprints Among Medical Academics: Convergent Mixed Methods Study"

Mustafa Sevim[1*], Dr med; Burak Karamese[2*]; Zafer Alparslan[2*]

[1]Department of Physiology, School of Medicine, Marmara University, Başıbüyük Yolu No: 9 D:2, Istanbul, Turkey

[2]School of Medicine, Marmara University, İstanbul, Turkey

[*]all authors contributed equally

Corresponding Author:
Mustafa Sevim, Dr med
Department of Physiology, School of Medicine, Marmara University, Başıbüyük Yolu No: 9 D:2, Istanbul, Turkey

Related Articles:

Companion article: https://www.medrxiv.org/content/10.1101/2025.04.09.25325524v1

Companion article: https://med.jmirx.org/2026/1/e95736

Companion article: https://med.jmirx.org/2026/1/e95737

Companion article: https://med.jmirx.org/2026/1/e78139

(JMIRx Med 2026;7:e95735) doi:10.2196/95735

KEYWORDS

preprint; medical academics; publishing attitudes; editorial policies; survey

This is the authors' response to peer review reports for "Awareness, Experiences, and Attitudes Toward Preprints Among Medical Academics: Convergent Mixed Methods Study."

Round 1 Review

Reviewer G [1]

1. Abstract: The authors [2] should specify the name of the university referred to as "a major university in Istanbul."

5. Methods: The authors should specify the name of the university referred to as "a major medical university in Istanbul." This information is important for assessing the reliability of the study and for confirming ethical approval in a transparent manner.

Response: The name of the university has been clearly placed where necessary.

2. Abstract: The authors should clarify how the responding editors and the biomedical journals that were manually reviewed were selected.

Response: Detailed information regarding journal policy review was covered in the "Editorial Perspectives from Turkish Biomedical Journals" and "Journal Policy Review" subsections of the "Methods" section. To better clarify this process, we added further information to related section and abstract and uploaded the whole journal list including biomedical journal categorization as a Multimedia Appendix file.

> *…Additionally, all responses to open-ended questions from journal editors and 118 biomedical journals were manually reviewed for their stated stance on preprints and article processing charges (APCs).*
>
> *…The email was sent to the editors-in-chief of all biomedical journals, and a total of 7 editors responded…*

3. Abstract: The authors are encouraged to present concrete data rather than relying solely on descriptive summaries.

Response: Additional concrete data have been added to the abstract.

> *...Subgroup analysis revealed that older participants scored higher on the "Preprint Test" (mean 2.20, SD 1.31 vs mean 1.97, SD 1.60) and had more experience with preprint publishing (2.5% of younger participants; 24.1% of older participants). Further, younger academics expressed less openness toward future use (17.5% in the younger group; 27.6% in the older group)...*

4. Introduction: The authors describe the study as "mixed method," but it lacks a legitimate integration process between the qualitative and quantitative components. Therefore, the term

"mixed method" should be avoided unless such integration is clearly demonstrated.

10. Methods: The section lacks a description of statistical analysis, making the analytical process unclear and only partially reported.

15. Results: The process for analyzing the qualitative responses is not clearly described, and the presentation of the results is extremely limited.

16. Results: The authors should provide information on how the responses from the editors were summarized. Without this explanation, reviewers and potential readers may find it difficult to interpret the results presented.

Response: Following these comments, the subheading "Study Design and Data Analyses" was added to the "Methods" section of the manuscript.

> *Study Design and Data Analyses*
>
> *This study employed a convergent mixed methods design integrating quantitative survey data, qualitative insights, and document-based content analysis to explore medical academics' awareness and attitudes toward preprints.*
>
> *A cross-sectional online questionnaire included demographic questions, Likert scales, and multiple-choice items. Descriptive statistics (frequencies, percentages, means, medians, SDs) were used and no inferential statistical testing was conducted. This survey was reported in accordance with the Checklist for Reporting Results of Internet E-Surveys (CHERRIES), and the completed checklist has been uploaded as a Multimedia Appendix file.*
>
> *Quantitative comparisons of journal policies were made using frequency counts and visualized via bar plots and heatmaps, with reference to impact metrics (eg, JCI quartiles).*
>
> *Open-ended responses within the survey were analyzed using pattern-based thematic analysis. Commonly expressed concerns were coded inductively to identify recurrent barriers and perceptions regarding preprint use. Responses were grouped into themes such as plagiarism concerns, lack of academic recognition, policy confusion, and ethical ambiguity.*
>
> *Editorial perspectives were obtained through open-ended email queries sent to biomedical journal editors. These responses were descriptively summarized to illustrate common institutional views and infrastructure limitations regarding preprint adoption.*
>
> *Findings from the three data sources were integrated during interpretation to identify convergence and divergence. Quantitative trends were contextualized with qualitative themes and policy landscape shifts, enabling a holistic understanding of both individual attitudes and institutional structures shaping preprint practices in Türkiye.*

6. Methods: The authors should indicate the total number of potential participants (ie, the total number of faculty members invited or eligible to participate).

Response: The total number of medical academics, which is the targeted population, was given in the "Participant Recruitment and Data Collection" subsection of the "Methods" section.

7. Methods: The authors should explain the rationale for dividing the age groups at 40 years.

Response: The mean age of our participants was 39.56, where median was 36 and ranging through 23 - 73. Moreover, age 40 can be considered as a turning point for a researcher in Türkiye. This age nearly corresponds to the time period when a researcher becomes an assistant professor. Altogether, we hypothesized that being above or below age 40 means something worth considering in terms of looking to science practice, which may affect attitude toward preprints. The necessary explanation on this subject has also been added to the "Subgroup Analyses" subsection of the "Methods" section.

> *Age: Participants were divided into two groups based on age; those younger than 40 and those 40 or older. This 40-year threshold was chosen for two primary reasons. First, it closely reflects the central tendency of our sample's age distribution (mean 39.56, median 36, range 23 - 73). Second, within the Turkish academic context, the age of 40 is a significant career milestone, often coinciding with the transition to an assistant professorship.*

8. Methods: The authors are encouraged to classify the biomedical journals into basic and clinical categories, in the same way that they categorized the survey respondents, even if some journals may cover both areas.

Response: Categorizing the journals as either "basic science" or "clinical science" was considered at the beginning. However, upon careful evaluation, it was concluded that this distinction could not be reliably applied. Since a significant majority of journals published content covering both areas, making a clear distinction was problematic. This ambiguity was compounded by the fact that the journals themselves do not use this classification system. Therefore, to maintain methodological rigor and avoid introducing subjective bias, no categorization was implemented. For full transparency, the complete list of journals is provided in a Multimedia Appendix file, should readers wish to perform their own classification.

9. Methods: The authors should provide a list of the journals included in this study.

Response: The journal list was uploaded as a Multimedia Appendix file and necessary citations were made in the relevant lines within the text.

11. Results: The authors should ensure that the findings are presented in alignment with the methods described. As the study does not involve a systematic review but rather a journal policy review, the current framing of the Results section may give a misleading impression.

Response: As noted by the reviewer, the term "systematic review" was avoided in the relevant sections to avoid creating a misleading impression.

12. Results: For clarity and coherence, the Results section should be reorganized to reflect the sequence of the study components—for example, starting with the questionnaire survey results, followed by the findings from the editorial and journal policy survey.

Response: We thank the reviewer for this thoughtful suggestion regarding the organization of the Results section. We agree that aligning the Results with the Methods is a standard and often effective approach. Therefore, we changed the heading "Editorial Perspectives" in the Methods section to "Editorial Perspectives from Turkish Biomedical Journals." In the "Results" section, we changed the heading "Preprint and APC Policies of Turkish Biomedical Journals" to "Journal Policy Review" and "Participant Demographics" to "Results of the Survey."

However, the order of the "Results" section was deliberately structured thematically to create a more logical and impactful narrative for the reader. The intention here is to first establish the broader context by presenting the findings from the editorial and journal policies. We believe this landscape is essential for the reader to fully understand and interpret the significance of the individual researchers' experiences detailed in the subsequent questionnaire survey results. This "macro-to-micro" progression strengthens the overall argument.

Therefore, after careful reconsideration of the reviewer's point, we have respectfully retained the original order, as we are confident it best serves the clarity and coherence of our findings.

13. Results: The authors mention the impact factor, but it is not described in the Methods section. Furthermore, the year and whether it represents the 2-year or 5-year impact factor are not specified.

Response: Only the 2-year journal impact factor is used throughout this study. The 2-year journal impact factor and journal citation index (JCI) quartiles are clarified in the "Editorial Perspectives from Turkish Biomedical Journals" subsection of the "Methods" section.

> *From an initial list of 280 journals, 264 remained after excluding duplicates, inaccessible websites, and journals with unclear policies. The 2-year impact factors (IF) and Journal Citation Index (JCI) quartiles were obtained as well.*

14. Results: The description of the preprint test, including its content and scoring method, is insufficient, making it difficult to assess its appropriateness.

Response: The meaning of "preprint test" and how this score was generated are further detailed under the subheading "Survey Instrument" in the "Methods" section. The full survey form is also included as a Multimedia Appendix file.

> *To quantify objective knowledge of preprints, a "preprint test score" was generated from 4 multiple-choice questions. Participants received 1 point for each correct response, resulting in a total score ranging from 0 to 4. Higher scores indicate higher knowledge of preprints. The answers given to the relevant part of the survey (9th question: "Tick the option you think is correct.") were used to calculate the "preprint test score." The whole survey form may be found in Multimedia Appendix 1.*

17. Discussion: The authors should revise the manuscript for logical consistency and explicitly discuss the limitations of this study prior to submitting it to another journal.

Response: The manuscript has been revised and the limitations section expanded as follows:

> *It is important to acknowledge that this study was conducted at a single academic institution in İstanbul. As such, our findings represent a localized snapshot and should be interpreted with caution. Additionally, while the survey captured a range of perspectives, the response rate was modest, and some questions had incomplete responses. The number of editor responses was also limited, restricting the depth of qualitative analysis. Finally, the policy review was limited to publicly available information, which may not fully reflect internal editorial practices or unpublished updates. Broader, multicenter studies will be necessary to determine whether these patterns hold across other regions and institutions in Türkiye.*

18. Tables: The authors should ensure consistent use of commas, periods, and digit formatting. Furthermore, the tables contain typographical errors that need correction.

19. References: The authors should review the reference formatting and ensure that it adheres to the journal's prescribed style.

Response: The format of the manuscript, including references and tables, was revised and edited based on the editor's and reviewer's comments.

Reviewer I [3]

This paper offers valuable insights into attitudes toward preprinting at a medical institution in Istanbul. The research appears sound to me. The paper is clear and easy to read.

I have one important comment: I was unable to find the survey form. I urge the authors to make the survey form, and also the data collected in the survey, openly available. To interpret the results of the survey, it is important to have access to the underlying survey questions.

Response: To ensure full interpretation of the results, the entire survey form is included as a Multimedia Appendix file.

References

1. Ide K. Peer review of "Awareness, Experiences, and Attitudes Toward Preprints Among Medical Academics: Convergent Mixed Methods Study". JMIRx Med 2026;7:e95736. [doi: 10.2196/95736]
2. Sevim M, Karamese B, Alparslan Z. Awareness, experiences, and attitudes toward preprints among medical academics: convergent mixed methods study. JMIRx Med 2025;7:e78139. [doi: 10.2196/78139]
3. Waltman L. Peer review of "Awareness, Experiences, and Attitudes Toward Preprints Among Medical Academics: Convergent Mixed Methods Study". JMIRx Med 2026;7:e95737. [doi: 10.2196/95737]

Edited by S Amal; submitted 19.Mar.2026; this is a non–peer-reviewed article;accepted 19.Mar.2026; published 17.Apr.2026.

Please cite as:
Sevim M, Karamese B, Alparslan Z
Authors' Response to Peer Reviews of "Awareness, Experiences, and Attitudes Toward Preprints Among Medical Academics: Convergent Mixed Methods Study"
JMIRx Med 2026;7:e95735
URL: https://xmed.jmir.org/2026/1/e95735
doi:10.2196/95735

Authors' Response to Peer Reviews of "Automating Individualized Notification of Drug Recalls to Patients: Complex Challenges and Qualitative Evaluation"

Meghana Gadgil[1], MD, MPH; Rose Pavlakos[2], PharmD; Simona Carini[1], MA; Brian Turner[3], MBA; Ileana Elder[4], PhD; William Hess[4], BS; Lisa Houle[5], BA; Lavonia Huff[4]; Elaine Johanson[4], BS; Carole Ramos-Izquierdo[4], MS, MPM; Daphne Liang[4], PharmD; Pamela Ogonowski[4], MLS (ASCP); Joshua Phipps[6]; Tyler Peryea[4], BA; Ida Sim[1,3], MD, PhD

[1]Division of General Internal Medicine, University of California, San Francisco, Box 0320, San Francisco, CA, United States

[2]Division of Cardiology, University of California, San Francisco, San Francisco, CA, United States

[3]Clinical and Translational Science Institute, University of California, San Francisco, San Francisco, CA, United States

[4]Food and Drug Administration, Silver Spring, MD, United States

[5]Tuvli, LLC, Herndon, VA, United States

[6]Conceptant, Inc, Falls Church, VA, United States

Corresponding Author:
Simona Carini, MA
Division of General Internal Medicine, University of California, San Francisco, Box 0320, San Francisco, CA, United States

Related Articles:

Companion article: https://www.medrxiv.org/content/10.1101/2024.09.18.24312141v1

Companion article: https://med.jmirx.org/2026/1/e82612

Companion article: https://med.jmirx.org/2026/1/e82613

Companion article: https://med.jmirx.org/2026/1/e68345

Abstract

(JMIRx Med 2026;7:e82609) doi:10.2196/82609

KEYWORDS

notification system; drug recalls; patient safety; medication; electronic health records; prescriptions; decision support

This is the authors' response to peer review reports for "Automating Individualized Notification of Drug Recalls to Patients: Complex Challenges and Qualitative Evaluation."

Round 1 Review

Reviewer E [1]

General Comments

This paper [2] describes a qualitative study that aims to leverage the US Food and Drug Administration's (FDA's) Healthy Citizen prototype platform, which provides information about recalls, to automatically notify patients of relevant recalls.

Specific Comments

Major Comments

1. Because of the setup of this document, it is challenging to be able to add comments or do any editing. Not sure what happened, but it treated every line as a single object when opened in Microsoft Word. Please check your formatting.

Response: We're sorry reviewing the document was difficult; we uploaded the document following the instructions provided. We hope the problem will not present itself again in the revised document uploaded after the review.

2. On page 2, within the abstract, under Background, there is an error in the formatting. There should be a section that begins with Aim. Instead, that section is folded into the Background section and needs to be corrected.

Response: Thank you for the comment. This has been corrected in the revised manuscript.

3. On page 8, with the MyChart message, I can see why patients felt too much wording was in this layout. Surprisingly, the Patient Advisory Council agreed to this layout and the wordiness. The focus must be on the patient's needs, not what the FDA requires. We all have seen the Prescribers' Digital Reference, and we know that the information is too dense and too small. This is similar to that in terms of format. Enlarge the font, eliminate extraneous information, and only include information that is important to the patient and in simple English. This should be pretty feasible in the formatting of the Healthy Citizen and/or the MyChart message.

Response: Thank you for the comment. We agree that clear and concise information is key in any communication with the patient. The design balanced FDA requirements, the information available on the Healthy Citizens platform, and the need to be accurate. As for the font size, please consider that the figure is an artifact, as the original screenshot needed to be shrunk to fit onto the manuscript page.

4. You identified problems and that patients would feel obligated to contact their provider regarding the recall. Instead of exploring how to address this so that patients wouldn't do that, thereby increasing the significant workload on the provider's health care team, you simply gave up. I think you could have done much more with this than say, "oh, it can't be done." How could you word the MyChart to direct the patients only to the pharmacy that dispenses their medication instead of the primary care provider? If you didn't ask that question, you should have. This is not the time to give up. It's time to inquire more to find the right answers so that this could move forward and better serve both the patients and their providers.

Response: Thank you for the comment. The MyChart message refers the patient only to the dispensing pharmacy and never mentions the prescribing physician or the clinic. However, during the qualitative evaluation, it became clear that the patients still wanted to discuss the recall with their clinicians. We felt that stronger wording, something along the lines of "Please do not contact your physician or the clinic on this matter," would have been detrimental and turned patients off.

5. It is certainly possible, given the technical requirements to create this capability, that you ran out of time and money. However, you can still benefit your team and others by focusing on the lessons learned and how you would go forward with another study.

Response: Thank you for your comment. As detailed in the paper, recall alerts sent to patients are not precise, but contacting the right patient for the appropriate recall is of paramount importance to avoid unnecessary anxiety and, worse, treatment discontinuation. False positives and the fact that patients expect their prescriber to be aware of, and involved in, responding to a drug recall, while prescribers don't have easy access to the relevant information, create an obstacle that another study would ultimately encounter and currently not be able to solve.

6. One of the things that you did not do is a first round of qualitative testing and using that feedback to make changes and do a second round. Per Nielsen [3], you only need about 5 test subjects per round to get the desired, usable results. What was preventing you from doing that? Put that in the manuscript as a limitation in your Discussion.

Response: Thank you for your comment. The study was planned based, among other things, on a project timeline. See the answer to comment 5 above.

7. Also, on page 11, in the last paragraph of the page under Discussion, there is a comment regarding patients expecting their providers to know when a recall has occurred; I think we all know this is an unreasonable expectation. Part of the communication with the MyChart message is to inform the patients not to call their provider but to call the pharmacy that dispenses their medication, which should be right on the bottle. Again, one component of the MyChart portal messaging system, as well as any other portal messaging system, is to keep patients informed and educate them. That should be a focus of this project, just as much as the technical components.

Response: Thank you for your comment. As mentioned in our answer to comment 4 above, we feel that a stronger wording, something along the lines of "Please do not contact your physician or the clinic on this matter," would have been detrimental and turned patients off.

8. On page 12, in the last full paragraph on the page, you make a statement regarding the project that a strong case can be made for requiring each pill bottle to include the lot number (maybe) and National Drug Code (NDC) of the pills. Since the FDA was a component of this project, that should probably have been something you recommended for the FDA to require and not leave to the state boards of pharmacy, as then you would get a patchwork of regulations. This would require the FDA to say that lot numbers and NDCs are required on the bottles of all medications with an appropriate implementation period to allow for appropriate software and hardware adjustments. That is just as valuable a recommendation out of the study as any other.

Response: Thank you for your comment. The FDA does not have the legal authority to regulate the practice of pharmacy in any state, and therefore the FDA cannot require that the lot number and NDC (or anything else, including the name of the drug) be placed on each prescription that a pharmacist dispenses to a patient. We clarified this in the Discussion section of the revised manuscript.

Reviewer F [4]

General Comments

This manuscript [2] describes interesting and novel work with far-reaching patient safety implications. The authors developed an automated system in the electronic health record (EHR) of an academic medical center that scans for drug recalls, matches up NDCs of recalled medication on a patient's medical list, and sends notifications through the EHR portal to the patient, providing them with more information on the recall. The authors then conducted a qualitative analysis of 9 patients' perceptions of a fictious recall notice. Despite successful development of the automated system, many limitations prevented the widescale

adoption of this system in 2 clinics associated with the large academic medical center. The outcome of the work—a decision was made not to deploy the new software for drug recalls—was surprising, and it is important that "failed" implementation work also be published. That said, key weaknesses of the manuscript are the lack of important details, need for better organization of the content, and the need for much stronger scientific and technical writing to accurately interpret the methods, results, and implications. These weaknesses also made it much more difficult to read and evaluate the manuscript. Despite the importance of the topic, the small sample size of patients also limits the work's impact.

Specific Comments

Title

1. It would be helpful if the title were a bit more specific about the technology, study methods (qualitative), and notification recipients (patients, providers, etc).

Response: Thank you for the suggestion. We edited the title to Automating Individualized Notification of Drug Recalls to Patients: Complex Challenges and Qualitative Evaluation."

Abstract

1. The Background section appears to be contradictory. Sentence 2 says the FDA has ways to notify health care professionals (HCPs) and patients, but then the following sentences seem to say the opposite.

Response: Thank you for your comment. The website referenced in the paper, "Recalls, Market Withdrawals, & Safety Alerts" [5], provides information to the public about recalls, but it does not notify HCPs about individual patients in their care who may be affected.

2. A few more details here on the type of platform would be helpful...software app? Web-based platform, etc? And what are the intended user types? (HCPs and patients? Or just patients?)

Response: Thank you for your comment. We added some details to the abstract.

3. The choice of methods doesn't seem to follow the Background section. Why was it necessary to include the clinics, rather than just work directly with the patients? Or, why was the focus on clinics, rather than pharmacies? (These comments apply to the main Introduction and Methods sections, as well.)

Response: Thank you for your question. As the study was implemented at an academic institution, we followed the institution's rules for engaging with patients.

4. I expected the "program description" to appear in the Methods section, not the Results.

Response: Thank you for your comment. The section was moved as suggested.

Introduction

1. The second and third sentences of the first paragraph of the Introduction: any studies or references to back up this claim?

Response: Thank you for your question. The paper's authors, who are HCPs, have extensive experience managing drug recalls in their daily practice. Published studies have focused on analyzing and classifying the recalls themselves (eg, the reason for the recall, the class).

2. No information is included on if/what literature explores this or similar topics.

Response: Please see the response to comment 1 under Introduction above.

3. I would recommend adding more information on the process pharmacies currently have in place for notifying patients of recalls. Also add any literature that exists showing how often patients then contact their providers or add quantitative data to highlight this extra burden on providers to emphasize the problem.

Response: Thank you for your comment. The process pharmacies follow is explained in the Drug Recall Process section. We found no literature on the topic.

4. I expected the funding information in the last sentence of the first paragraph to be included in a funding statement or the acknowledgments (rather than the Introduction) and the rest of that statement to be described in the Methods.

Response: Thank you for your comment. We moved the paragraph as suggested.

Setting

1. I expected this to appear under a larger Methods section.

Response: We added a Methods section.

2. What was the goal sample size and rationale for the sample size? There is missing demographic information on the participating patients.

Response: A convenience sample was used based on outreach to patients and their responses. Given the nature of the study, we feel we provided sufficient information on the patients interviewed.

3. So the Fast Healthcare Interoperability Resources (FHIR) portion notified HCPs? The intended recipients are not specified for that part of the program.

Response: The notification was meant for patients only.

4. "EHR build" was unexpected as a reader. Is that a third part? How does it fit into the first 2 parts?

Response: We added a clarification.

5. The screenshots and figures are useful.

Response: Thank you for your comment. We are glad the figures are useful.

6. Even for a convenience sample, more details are needed on recruitment. How did you choose which patients to email? How many were emailed for recruitment? Were patients emailed and recruited sequentially, for example? Were there any exclusion or inclusion criteria for patients? Did any patients decline to participate? Why? What was the distribution of patients recruited from primary care versus cardiology?

Response: Recruitment techniques were not a subject of this study. Patient inclusion was based on active use of the MyChart portal.

7. More specific details are warranted for the methods used for qualitative analysis, such as whether an inductive versus a deductive design was used. Was a consensus approach used, or some other approach? See also the writing guidelines for qualitative studies (eg, the Consolidated Criteria for Reporting Qualitative Research [COREQ], Standards for Reporting Qualitative Research [SRQR]). Explain also the "additional verification" process during analysis. References should be cited for the qualitative methods used in this work.

Response: We added the interview script as an appendix. Analysis of the responses was repeated and the results compared.

8. Did any of the patient sample have prior experience with MyChart, and if so, what was the average number of years of MyChart experience?

Response: Active use of MyChart was an inclusion criterion.

9. These statements from the text appear to be contradictory, and the meaning of the first statement especially is unclear, and seems like an opinion: "[Patients expressed that the] widget should not ask patients to discuss the information with their healthcare provider." "Patients wanted to discuss the recall with their clinicians to 'close the loop.'"

Response: Figure 2 shows the information provided by the FDA's Healthy Citizen platform, which complies with the FDA's requirements and is not customizable by the system using it.

10. The conclusion not to deploy the system seems dramatic based on the findings and makes me wonder if any other creative solutions were considered to address the concern of potential increased clinic burden. Also, how was it determined that the clinic burden outweighed safety risks to the patient? Maybe the system should only be used for certain types of recalls, for example. Or maybe the system could be integrated more with the pharmacy, rather than the prescriber's clinic, or the letter could read differently (advising against contacting the clinic unless the patient was unable to resolve the issue with the pharmacy). Or the letter could explain that only the pharmacy, not the clinic, would have a record of the patient's specific manufacturer and whether the recall applied to them.

Response: The MyChart message explains that the pharmacy has more information about the drug given to the patient; the message cannot state for certain that the pharmacy can match the recall precisely to the patient.

11. It would be helpful to see the full interview guide and patient scenario details in a supplementary appendix to aid interpretation of the methods and results.

Response: We added the interview script as a multimedia appendix.

Discussion

1. The Discussion does not mention limitations of the study design and methods.

Response: Our project was technically successful. The lack of availability of the data needed to accurately target patients—particularly the lot number of the drug dispensed—makes false positive notifications unavoidable, independent of the study design.

3. Is anything stamped on the medication (eg, pill) itself to indicate the manufacturer? Or is that also inconsistent across medications?

Response: As described in the Discussion section, what data pertaining to the drug appears where is not consistent. (And while the pharmacy records the NDC of filled prescriptions, pills from different lot numbers can be dispensed together.)

5. In the last paragraph of the Discussion, there is no citation for the number of state boards of pharmacy that require the lot number to appear on the label.

Response: While we understand the interest in learning which state boards of pharmacy require the lot number to appear on the label, considering that the number is one-tenth of the states, the takeaway is that the problem described applies to the vast majority of the states.

6. I expected the Discussion to close with a Conclusions paragraph outlining key lessons learned and any generalizable findings.

Response: In the Conclusions we reiterate the need for consistent availability of the data needed to accurately address patients affected by a drug recall.

Round 2 Review

Reviewer E

General Comments

This paper describes a small qualitative study that aims to leverage the FDA's Healthy Citizen prototype platform, which provides information about recalls, to notify patients of relevant recalls automatically. The project team deemed the goal unattainable and provided limited lessons learned and recommendations for potential advocacy/future solutions.

Specific Comments

Major Comments

1. On page 11, in the section/paragraph beginning with "Major thematic findings included...": these are some of the lessons learned that I mentioned in my feedback.

Response: Thank you for your suggestion. We added content to the Conclusions paragraph summarizing lessons learned and outlining the generalizability of some of them.

2. On page 12, in the paragraph beginning with "The project team concluded that...": The "project team" felt this. Did the Patient Advisory Council and the test subjects share the same feeling?

Response: Thank you for your question. We did not go back to the Patient Advisory Council or to the test subjects after the conclusion of the project. While their support for expanding the

pilot would have been encouraging, their support could not solve the challenges encountered during the project implementation. An expansion would have required institutional support. While the project implementation provided important lessons, it did not provide a solid enough business case to justify expanding the pilot. We added this to the Program Evaluation section.

3. On page 13, in the second paragraph on the page, in the sentence beginning with "Note that the FDA does not...": this would clearly be a lesson learned and could be advocated for via Congress and the Department of Health and Human Services.

Response: Thank you for your comment. We added the following content to the Conclusions paragraph to address it: "While a change at the federal level would be ideal, advocating individual State Boards of Pharmacy to require the NDC and lot number to appear on the dispensed medication label may provide interim needed progress allowing development and deployment of solutions supporting patients' needs."

4. On page 13, in the the second paragraph, the next sentence, beginning with "The manufacturer and lot number of dispensed medications...": agreed. See previous comment.

Response: Thank you for your comment. We added content to the Conclusions paragraph to address it. See our response to comment 3 above.

Reviewer F

General Comments

The authors addressed a few of my review comments and made some text changes, but unfortunately, most of my comments—about 15 of them—remain inadequately addressed. For the comments listed again below, the authors did not appear to change anything in the manuscript to address the comment. In many cases, even the authors' reply to the reviewers did not answer the question. Also, the authors describe adding the interview guide as an appendix, but I could not find this file on the reviewer website.

Response: Regarding the interview guide: the file was uploaded on April 12 on the authors' submission website, ahead of the resubmission and recirculation of the manuscript. Right-clicking on the file name identifies the URL [6].

Specific Comments

Abstract

1. The Background section appears to be contradictory. Sentence 2 says the FDA has ways to notify health care professionals (HCPs) and patients, but then the following sentences seem to say the opposite.

Response: The FDA has public-facing resources, including the Recalls, Market Withdrawals, & Safety Alerts website [5], which can be consulted by anyone. However, as mentioned in the Abstract, prescribers are not notified individually and specifically about which of their patients are affected by a recall. We added some words to clarify the distinction between general and specific and deleted the last sentence.

3. The choice of methods doesn't seem to follow the Background section. Why was it necessary to include the clinics, rather than just work directly with the patients? Or, why was the focus on clinics, rather than pharmacies? (These comments apply to the main Introduction and Methods sections, as well.)

Response: The project's premise was that patients seek answers to recall-related questions from their HCPs. Therefore, we wished to answer the question at the levels of primary care and a cardiology clinic. We worked with the project principal investigators' clinics and patients and did so following the applicable requirements.

Introduction

2. No information is included on if/what literature explores this or similar topics.

Response: We added 4 references, 3 of them to recently published papers focused on the analysis of recall-related data (see the response to question 2 under Discussion below for summary details]

Setting

2. What was the goal sample size and rationale for the sample size? There is missing demographic information on the participating patients.

Response: As previously noted, the convenience sample was based on outreach to patients and their responses. Given the exploratory nature of the study, we feel we provided sufficient information on the patients interviewed. Power calculation and balancing the sample for certain variables were not relevant.

3. So the FHIR portion notified HCPs? The intended recipients are not specified for that part of the program.

Response: Thank you for your question. No, the HCPs did not receive any notification. The Healthy Citizens (SMART-on-FHIR) widget was launched from the MyChart message sent to the patient. We added a sentence between Figures 1 and 2 to clarify.

6. Even for a convenience sample, more details are needed on recruitment. How did you choose which patients to email? How many were emailed for recruitment? Were patients emailed and recruited sequentially, for example? Were there any exclusion or inclusion criteria for patients? Did any patients decline to participate? Why? What was the distribution of patients recruited from primary care versus cardiology?

Response: Thank you for your questions. We added some details to the manuscript in response. Established patients at the Department of General Internal Medicine (primary care) clinic who were members of the Patient Advisory Council, used MyChart, and were prescribed at least one medication received a recruitment letter. Patients at the cardiology clinic who were scheduled to see the pharmacist during a random week, who actively used MyChart (or their family members who used MyChart on their behalf), and who used at least one prescription medication were deemed eligible for the study and sent a recruitment letter. Interested patients contacted the study team to participate. Nine patients were interviewed.

7. More specific details are warranted for the methods used for qualitative analysis, such as whether an inductive versus a deductive design was used. Was a consensus approach used, or some other approach? See also the writing guidelines for qualitative studies (eg, the COREQ, SRQR). Explain also the "additional verification" process during analysis. References should be cited for the qualitative methods used in this work.

Response: Thank you for your question. The objective of the interviews was to obtain qualitative feedback from patients and identify the feedback's main themes using a consensus approach (a reference has been added to the manuscript). As detailed in the manuscript, the recordings of the interviews were transcribed and separately analyzed by 2 investigators to identify common themes, then 2 other team members verified the initial analysis. These themes are described in the manuscript in the paragraph starting with "Major thematic findings included the following..."

8. Did any of the patient sample have prior experience with MyChart, and if so, what was the average number of years of MyChart experience?

Response: The 9 patients interviewed were all MyChart users. We clarified in the manuscript that MyChart use was an inclusion criterion. We did not consider the number of years of MyChart experience as a relevant data point.

9. These statements from the text appear to be contradictory, and the meaning of the first statement especially is unclear, and seems like an opinion: "[Patients expressed that the] widget should not ask patients to discuss the information with their healthcare provider." "Patients wanted to discuss the recall with their clinicians to 'close the loop.'"

Response: The suggestion to discuss the recall information with the health practitioner was displayed on the FDA Health Citizen widget and could not be modified. We clarified this in the manuscript. The interviews confirmed that the statement led to confusion. The MyChart message recommended calling the pharmacy, as it would be the entity with more information to help the patient verify whether the recall applied to them (Figure 1).

10. The conclusion not to deploy the system seems dramatic based on the findings and makes me wonder if any other creative solutions were considered to address the concern of potential increased clinic burden. Also, how was it determined that the clinic burden outweighed safety risks to the patient? Maybe the system should only be used for certain types of recalls, for example. Or maybe the system could be integrated more with the pharmacy, rather than the prescriber's clinic, or the letter could read differently (advising against contacting the clinic unless the patient was unable to resolve the issue with the pharmacy). Or the letter could explain that only the pharmacy, not the clinic, would have a record of the patient's specific manufacturer and whether the recall applied to them.

Response: The MyChart message recommended calling the pharmacy as it is the entity with more information to help the patient verify whether the recall applied to them (Figure 1). Patients contacting the clinic received the same instructions. Most pharmacies have protocols in place to handle recalls, which may include outreach to customers. Integration with pharmacies was out of scope for this project and would have been a substantial undertaking: just the 2 clinics involved in the project serve over 37,000 patients, who fill their prescriptions in different pharmacies, from large chains to small local pharmacies to online ones. In the manuscript, we mention integrating with Surescript via claims data. However, such integration would not cover all the institution's patients, and Surescript records do not include dispensed lot numbers, so the problem of false positive notification would still exist. Should funding become available, we do not rule out exploring alternative solutions in the future. In response to a comment from the other reviewer, we added in the Program Evaluation section that while the project implementation provided important lessons, it did not provide a solid enough business case to justify expanding the pilot, which would have required institutional support.

Discussion

1. The Discussion does not mention limitations of the study design and methods.

Response: The project did not move forward for reasons that go beyond the qualitative evaluation we performed (see also our response to comment 10 under Setting, above).

2. I expected at least some comparison to other, related literature.

Response: We added references to recently published papers:

An analysis of FDA drug recall data (2012-2023) showing that drug recalls are frequent [7]. The paper talks about the causes of drug recalls and suggests improvements to the relevant FDA database, but it doesn't discuss the impact of recalls on clinical care.

A study of drug recalls in the Netherlands, which also identifies the issue that pharmacists do not always know which batch was dispensed to a patient [8].

An analysis of the clinical impact of the 2018 recalls of several angiotensin II receptor blockers and the impact in terms of medication gap and clinical outcomes [9].

These are recently published supporting articles that analyze existing data. None includes a program such as ours.

3. Is anything stamped on the medication (eg, pill) itself to indicate the manufacturer? Or is that also inconsistent across medications?

Response: What is printed on an individual solid oral-dosage-form product (eg, tablet or capsule) depends on the manufacturer complying with 21CFR206.10(a) in the Code of Federal Regulations [10]. In the United States, most solid oral-dosage-form drug products are required to have an imprint code (eg, logo, letters, numbers, or a combination). As detailed in the manuscript, at the federal level, the FDA does not have the legal authority to regulate the practice of pharmacy in any state and cannot require that specific information be placed on each prescription label that a pharmacist dispenses to a patient. Individual states (via their state boards of pharmacy) regulate what appears on the pill bottle label and on the leaflet provided to the patient alongside the medication.

4. A table of key recommendations could strengthen the paper.

Response: Thank you for your suggestion. We have added a list of lessons learned and a recommendation at the end.

5. In the last paragraph of the Discussion, there is no citation for the number of state boards of pharmacy that require the lot number to appear on the label.

Response: We added the requested details to the statement pertaining to lot number requirements and added the relevant supporting references. No peer-reviewed synthesis exists on this point, so we relied on primary legal sources. We also amended the original statement pertaining to the NDCs to clarify the rules and the issuing body: "As of August 2025, our review of state regulations identified the following jurisdictions with explicit requirements. Four State Boards of Pharmacy (Colorado, Delaware Oklahoma, Wyoming) plus the U.S. territory of Puerto Rico require the lot number to appear on the dispensed medication label [12-16]. The Pennsylvania State Board of Medicine requires the NDC to appear on the dispensed medication label if the prescriber specifies that the drug name not appear on the label [17]. The State Boards of Pharmacy of New Hampshire and Ohio, allow the use of NDC as abbreviation for the manufacturer / distributor name [18-19]."

Acknowledgments

The manuscript's contents are solely the responsibility of the authors and do not necessarily represent the official views of the US Department of Health and Human Services (HHS) or the FDA. Any policy recommendations in this manuscript are offered for discussion and do not represent HHS or FDA policy or commitments.

References

1. Marshall RC. Peer review of "Automating Individualized Notification of Drug Recalls to Patients: Complex Challenges and Qualitative Evaluation". JMIRx Med 2026;7:e82612. [doi: 10.2196/82612]
2. Gadgil M, Pavlakos R, Carini S, et al. Automating individualized notification of drug recalls to patients: complex challenges and qualitative evaluation. JMIRx Med 2026;7:e68345. [doi: 10.2196/68345]
3. Why you only need to test with 5 users. Nielsen Norman Group. URL: https://www.nngroup.com/articles/why-you-only-need-to-test-with-5-users/ [accessed 2025-10-14]
4. Russ A. Peer review of "Automating Individualized Notification of Drug Recalls to Patients: Complex Challenges and Qualitative Evaluation". JMIRx Med 2026;7:e82613. [doi: 10.2196/82613]
5. Recalls, market withdrawals, & safety alerts. Food and Drug Administration. 2020. URL: https://www.fda.gov/safety/recalls-market-withdrawals-safety-alerts [accessed 2025-10-14]
6. Testing protocol script. JMIR Publications. URL: https://xmed.jmir.org/api/download?filename=de17a001f50f97f6a02532759495f1c9.pdf&alt_name=68345-1130394-1-SP.pdf [accessed 2025-10-14]
7. Ghijs S, Wynendaele E, De Spiegeleer B. The continuing challenge of drug recalls: insights from a ten-year FDA data analysis. J Pharm Biomed Anal 2024 Oct 15;249:116349. [doi: 10.1016/j.jpba.2024.116349] [Medline: 39029352]
8. Annema PA, Derijks HJ, Bouvy ML, van Marum RJ. Impact of drug recalls on patients in the Netherlands: a 5‐year retrospective data analysis. Clin Pharma and Therapeutics 2024 Jun;115(6):1365-1371 [FREE Full text] [doi: 10.1002/cpt.3220]
9. Callaway Kim K, Roberts ET, Donohue JM, et al. Changes in blood pressure, medication adherence, and cardiovascular-related health care use associated with the 2018 angiotensin receptor blocker recalls and drug shortages among patients with hypertension. J Manag Care Spec Pharm 2025 May;31(5):461-471. [doi: 10.18553/jmcp.2025.31.5.461]
10. Title 21, Chapter |, Subchapter c, Part 206, § 206.10. Code of Federal Regulations. URL: https://www.ecfr.gov/current/title-21/chapter-I/subchapter-C/part-206/section-206.10 [accessed 2025-10-16]

Abbreviations

COREQ: Consolidated Criteria for Reporting Qualitative Research
EHR: electronic health record
FDA: Food and Drug Administration
FHIR: Fast Healthcare Interoperability Resources
HCP: health care professional
NDC: National Drug Code
SRQR: Standards for Reporting Qualitative Research

Edited by CN Hang; submitted 18.Aug.2025; this is a non–peer-reviewed article;accepted 18.Aug.2025; published 13.Jan.2026.

Please cite as:
Gadgil M, Pavlakos R, Carini S, Turner B, Elder I, Hess W, Houle L, Huff L, Johanson E, Ramos-Izquierdo C, Liang D, Ogonowski P, Phipps J, Peryea T, Sim I
Authors' Response to Peer Reviews of "Automating Individualized Notification of Drug Recalls to Patients: Complex Challenges and Qualitative Evaluation"
JMIRx Med 2026;7:e82609
URL: https://xmed.jmir.org/2026/1/e82609
doi:10.2196/82609

Authors' Response to Peer Reviews of "Administration Technique of Intranasal Corticosteroid Sprays Among Nepali Pharmacists: Cross-Sectional Study"

Amar Prashad Chaudhary[1], PharmD; Suraj Kumar Thakur[2], BPharm; Shiv Kumar Sah[2], BPharm, MPharm

[1]Tribhuvan University Teaching Hospital, Maharajgunj, Kathmandu, Nepal

[2]Institute of Medicine, Maharajgunj Medical Campus, Tribhuvan University, Kathmandu, Nepal

Corresponding Author:
Amar Prashad Chaudhary, PharmD
Tribhuvan University Teaching Hospital, Maharajgunj, Kathmandu, Nepal

Related Articles:

Companion article: http://preprints.jmir.org/preprint/83042

Companion article: https://med.jmirx.org/2026/1/e91439

Companion article: https://med.jmirx.org/2026/1/e91443

Companion article: https://med.jmirx.org/2026/1/e83042

(JMIRx Med 2026;7:e91445) doi:10.2196/91445

KEYWORDS

intranasal corticosteroid spray; allergic rhinitis; device use technique; pharmacist; patient counselling, continuing pharmacy education

This is the authors' response to peer-review reports for "Administration Technique of Intranasal Corticosteroid Sprays Among Nepali Pharmacists: Cross-Sectional Study."

Round 1 Review

Reviewer AH [1]

General Comments

This paper [2] addresses an important gap by evaluating pharmacists' proficiency in demonstrating intranasal corticosteroid technique, using a standardized 12-step checklist with 5 critical steps. The sample size (n=365) is reasonable for a local study, and the use of multivariate logistic regression and Chi-square automatic interaction detection decision tree analysis adds analytical depth. The findings highlight systemic issues, such as inadequate training and curriculum gaps, which could inform policy changes to improve allergic rhinitis management and reduce adverse effects like epistaxis.

Specific Comments

Major Comments

1. Simple random sampling was used for pharmacies, but details on how wards were selected or how pharmacists within pharmacies were approached are vague. Please supplement and elaborate on further details of the randomization. More information on such would help lower the selection bias (eg, busier or more accessible pharmacies might be overrepresented).

Response: Thank you for this valuable comment. We agree that additional clarification regarding the sampling process is important to demonstrate methodological rigor and minimize concerns about selection bias. In the revised manuscript, we have now expanded the description of the sampling procedure. We have clarified how the wards in Kathmandu district were included, how the list of pharmacies was prepared, how simple random sampling of pharmacies was actually conducted, how pharmacists within each selected pharmacy were approached, and how the accessibility or busyness of pharmacies was handled.

2. The questionnaire's validity is only face-validated by experts, with no content or construct validity testing mentioned. Reliability was assessed via Cronbach alpha (0.758) on a small pilot (n=15), which is acceptable but not robust. The cutoff for "adequate" proficiency (>6/12 marks) is based on the median score and expert opinion, which feels arbitrary and not clinically validated. Why not base it on critical steps alone, given their emphasis on efficacy and safety? Only 6% performed all 5 critical steps correctly, yet 47% were deemed "adequate" overall. This discrepancy suggests the threshold may be too lenient, masking true incompetence in high-impact areas like directing the nozzle away from the septum (to prevent epistaxis)

or exhaling through the mouth (to optimize deposition). Please address these in the Discussion section.

Response: Thank you for this comment. The 12-step intranasal corticosteroid checklist was developed from established international guidelines (eg, ARIA, Benninger et al [3], NHS), ensuring content relevance. As this tool assesses observed procedural technique, construct validity testing is not applicable. We have clarified this and acknowledged the limitation in the Discussion.

We agree the pilot sample was small. Cronbach alpha of 0.758 represents acceptable internal consistency for an observational checklist. The limitation has now been explicitly acknowledged.

The score of more than 6 is not arbitary; we conducted a sensitivity analysis using alternative cutoffs (>5 and >7). Receiver operating characteristic analysis could not be performed because the total score forms the derived outcome without an external gold standard. Sensitivity analysis showed that (1) predictors remained stable and significant at >5 and >6 and (2) the >7 cutoff produced unstable models due to sparse cell counts. Thus, the >6 threshold is empirically supported, aligns with >50% competency, the median distribution, and expert opinion. Relevant text was added to the Methods and Discussion.

Only 6% of participants completed all critical steps; using this as the cutoff would create extremely low event counts and make regression analysis unreliable. Moreover, international guidelines require all 12 steps for complete patient counseling. We expanded the Discussion to highlight the clinical significance of poor critical-step performance.

3. Self-reported variables (eg, counseling frequency, use of materials) are prone to recall or social desirability bias, especially in an in-person interview setting. Please supplement these in the Discussion section.

Response: Thank you for the concern. The risk of recall or social desirability bias is mentioned in the Discussion section.

4. The multivariate binary logistic regression identifies associations (eg, male gender, older age, higher qualifications linked to better proficiency), but potential confounders like pharmacy type (independent vs chain) or workload details are not controlled for. Odds ratios are extreme in places (eg, BPharm holders 97% less likely to perform inadequately, or frequent counselors 11 times more proficient), which may stem from small subgroups or multicollinearity.

Response: In Nepal, most of the pharmacies are independently owned and very few are chain pharmacies. In this study, only independent pharmacies were used, therefore pharmacy type is not one of the potential confounders in this study. However, workload details as potential confounders were not measured, which may have partly contributed to the large adjusted odds ratio of some predictors. It is mentioned in the Discussion.

5. Gender differences (males ~2 times more proficient) were found but underlying factors were not explored (eg, access to workshops, cultural biases). Please elaborate more or address the potential underlying factors in the Discussion section.

Response: We thank the reviewer for highlighting this point. Additional contextual explanation has been added to the Discussion to address potential underlying factors.

6. "Educational materials" are linked to better proficiency, but what constitutes these (eg, leaflets, videos)? Please specify for readers to enhance the proficiency on applying the study's results.

Response: Thank you for the concern. The term "educational materials" mean the leaflet and now it is clearly mentioned in the Results.

7. Reference 16 has the wrong format for the volume, issue, and page numbers:

Al-Taie A. A Systematic Review for Improper Application of Nasal Spray in Allergic Rhinitis: A Proposed Role of Community Pharmacist for Patient Education and Counseling in Practical Setting. Asia Pacific Allergy. 2025:10 - 5415.

The full information from PubMed is as below:

Al-Taie A. A systematic review for improper application of nasal spray in allergic rhinitis: A proposed role of community pharmacist for patient education and counseling in practical setting. Asia Pac Allergy. 2025 Mar;15(1):29 - 35. doi: 10.5415/apallergy.0000000000000173. Epub 2025 Jan 13. PMID: 40051424; PMCID: PMC11882221.

Therefore, "2025:10 - 5415" should be "2025 Mar;15(1):29 - 35."

Please revise the whole reference list to see if any other typos exist.

Response: Thank you for the concern. All the references have been revised.

Reviewer AL [4]

General Comments

This is an important and well-written study. My suggestions are listed below.

Specific Comments

There are some problems with language and with unnecessary capitalization of words.

Page 3: INCS sprays should be defined in full on first mention in the text.

Response: Thank you for the concern. INCS spray is defined in full on first mention.

Page 8: Can details of the ethical committee that provided the approval be provided? Was the informed consent obtained in writing?

Response: The ethical committee details are now added in the manuscript. Yes, written informed consent was obtained from the participants.

Scoring system: Should the crucial steps not be provided with greater marks compared to the other steps?

Response: We thank the reviewer for this insightful suggestion. Although the five steps marked as "critical" have a greater clinical impact on efficacy and safety, we deliberately assigned equal weight (1 mark per step) to all 12 steps to maintain consistency with previously published studies that used similar checklist-based scoring systems and to avoid introducing subjective weighting without formal validation.

To address the clinical importance of critical steps, we analyzed them separately and reported their performance independently. Notably, although 47.1% of pharmacists met the overall adequacy threshold, only 6% correctly demonstrated all five critical steps, highlighting a substantial gap that would have been masked even if weighted scoring were used.

We agree that weighted scoring systems may better reflect clinical risk; however, such systems require prior validation. We have therefore added this point to the Limitations section and recommend weighted or competency-based scoring models in future studies.

Page 17: Please explain the classification tree (Chi-square automatic interaction detection method) for the benefit of the readers.

Response: We thank the reviewer for this helpful suggestion. We have now added a brief explanation of the classification and regression tree analysis using the Chi-square automatic interaction detector method in the Statistical Analysis and Results sections. The revised text explains the purpose of the method, the basis of variable splitting, and how the resulting tree should be interpreted. This addition is intended to improve clarity and accessibility for readers who may be unfamiliar with decision tree–based methods.

Page 17: "This research is one of a kind, conducted in Nepal." Can this sentence be modified?

Response: Thank you for the insightful suggestion. The sentence has been modified.

Page 19: Instead of continuing medical education (CME), continuing pharmacy education (CPE) may be a better term.

Response: Thank you for the suggestion. Continuing pharmacy education (CPE) has been used instead of continuing medical education (CME) in the manuscript.

Page 20: What educational aids are you referring to?

Response: Educational aids means the leaflets and that has been clarified in the manuscript now.

Are the educational leaflets available in the Nepali language?

Response: The educational leaflet was available in the English language and the pharmacist used it for reference while counseling the patients.

Page 20: "In our study, both the increasing age (> 26 y old) were significantly associated with improved INCS [intranasal corticosteroid] counseling proficiency." This sentence mentions both but then highlights only one factor.

Response: Thank you for the suggestion. It was a typing error in the manuscript and has been corrected.

Was this study conducted only in Kathmandu city and not in Lalitpur or Bhaktapur?

Response: This study was extensively conducted only in Kathmandu district.

Page 21, Limitations section: Some of the findings may be extreme due to small subgroups or model overfitting. Can this be explained?

Response: Thank you for the suggestion. This has been explained in the Limitations section.

Different fonts are used in different locations, and this should be corrected.

Response: Thank you for the comment. Font size has been corrected.

References

1. Au SCL. Peer review for "Administration Technique of Intranasal Corticosteroid Sprays Among Nepali Pharmacists: Cross-Sectional Study". JMIRx Med 2026;7:e91439. [doi: 10.2196/91439]
2. Chaudhary AP, Thakur S, Sah SK. Administration technique of intranasal corticosteroid sprays among Nepali pharmacists: cross-sectional study. JMIRx Med 2026;7:e83042. [doi: 10.2196/preprints.83042]
3. Benninger MS, Hadley JA, Osguthorpe JD, et al. Techniques of intranasal steroid use. Otolaryngol Head Neck Surg 2004 Jan;130(1):5-24. [doi: 10.1016/S0194-5998(03)02085-0] [Medline: 14726906]
4. Shankar RP. Peer review for "Administration Technique of Intranasal Corticosteroid Sprays Among Nepali Pharmacists: Cross-Sectional Study". JMIRx Med 2026;7:e91443. [doi: 10.2196/91443]

Edited by A Schwartz; submitted 14.Jan.2026; this is a non–peer-reviewed article;accepted 14.Jan.2026; published 29.Jan.2026.

Please cite as:
Chaudhary AP, Thakur SK, Sah SK
Authors' Response to Peer Reviews of "Administration Technique of Intranasal Corticosteroid Sprays Among Nepali Pharmacists: Cross-Sectional Study"
JMIRx Med 2026;7:e91445
URL: https://xmed.jmir.org/2026/1/e91445
doi:10.2196/91445

Authors' Response to Peer Reviews of "The Performance of DeepSeek R1 and Gemini 3 in Complex Medical Scenarios: Comparative Study"

Maria Bajwa[1*], PhD; Robert Hoyt[2*], MD; Dacre Knight[3*], MD; Maruf Haider[4], MD

[1]MGH Institute of Health Professions, Boston, MA, United States

[2]Internal Medicine Department, Virginia Commonwealth University, 57 North 11th Street, Richmond, VA, United States

[3]Internal Medicine Department, University of Virginia, Charlottesville, VA, United States

[4]Internal Medicine Department, Carilion Roanoke Memorial Hospital, Roanoke, VA, United States

[*]these authors contributed equally

Corresponding Author:
Robert Hoyt, MD
Internal Medicine Department, Virginia Commonwealth University, 57 North 11th Street, Richmond, VA, United States

Related Articles:

Companion article: https://www.medrxiv.org/content/10.1101/2025.04.29.25326666v1

Companion article: https://med.jmirx.org/2026/1/e96223

Companion article: https://med.jmirx.org/2026/1/e96225

Companion article: https://med.jmirx.org/2026/1/e96227

Companion article: https://med.jmirx.org/2026/1/e76822

(JMIRx Med 2026;7:e96220) doi:10.2196/96220

KEYWORDS

large reasoning model; LRM; large language model; LLM; accuracy; medical scenario; DeepSeek R1; Gemini 3

This is the authors' response to the peer review of "The Performance of DeepSeek R1 and Gemini 3 in Complex Medical Scenarios: Comparative Study."

Round 1 Review

Reviewer B [1]

General Comments

This is a timely and well-structured paper [2] that investigates the application of DeepSeek R1, a state-of-the-art large reasoning model, in the medical domain using the Massive Multitask Language Understanding Pro (MMLU-Pro) benchmark. This paper presents a follow-up evaluation of the DeepSeek R1 large reasoning model on open-ended medical scenarios from the MMLU-Pro benchmark. The study finds that DeepSeek R1 achieves a high accuracy of 92% without multiple-choice options, demonstrating its potential utility in more realistic clinical settings. The paper is timely and relevant, with strong empirical results and clear motivation. However, it would benefit from revisions to improve clarity, contextual grounding in existing work, and methodological detail. The authors may also consider citing recent work that examines the questioning strategies of large language models (LLMs) in clinical dialogues to better position this study in the broader landscape.

Response: We appreciate the reviewer's comment, and we have now added citations and improved the clarity and contextual grounding of the paper.

The study is commendable for its effort in combining expert validation with benchmark testing and highlighting both performance and interpretability aspects. The paper is generally well-written, informative, and relevant to the research community on artificial intelligence (AI) in health care. However, before being suitable for publication, several important revisions are required. These include expanding the related work section to better situate the contribution of current research efforts, addressing some methodological limitations more transparently, and improving the robustness and generalizability.

Response: We thank the reviewer for this comment and supportive critique. We have edited our conclusions and mentioned generalizability in the Discussion section. We have

also combined the results of querying closed- and open-ended scenarios.

Major Comments

1. *While the paper references MMLU, MedQA, and some domain-specific LLM evaluations, it lacks a deeper discussion on recent approaches to questioning capabilities and long-context understanding in medical AI. Two notable papers should be included. First, "HealthQ: Unveiling Questioning Capabilities of LLM Chains in Healthcare Conversations" by Wang et al [3]. This paper presents a benchmarking framework focusing on the inquiry and elicitation capacity of LLM chains, which directly relates to the "reasoning" and prompt design aspects discussed here. Second, "Context Clues: Evaluating Long Context Models for Clinical Prediction Tasks on EHR Data" by Wornow et al [4]. This study highlights how context windows and task framing affect LLM performance on clinical reasoning—relevant for understanding how question complexity and format might interact with LLM accuracy. The paper could be strengthened by referencing more recent work on prompting and questioning strategies in clinical LLM applications. The paper would benefit from referencing Wang et al [3], which evaluates LLM chains' ability to optimize questions through reflection and prompting. This is relevant to the current paper's interest in open-ended diagnostic reasoning and LLM behavior in clinical settings.*

Response: We thank the reviewer for providing this information and for highlighting these notable papers. We have now accordingly mentioned the important work done by these researchers and also cited these papers in the Introduction section.

2. Although the DeepSeek R1 model is rigorously evaluated against MMLU-Pro, there's a lack of direct performance comparison to other LLMs (eg, MedPaLM, GPT-4, Claude) on the same dataset or medical scenarios. Even informal or partial benchmarks would help contextualize the model's effectiveness. Also, the novelty should be better emphasized—is this the first comprehensive large reasoning model evaluation on MMLU-Pro's health subset?

Response: This is the first comprehensive evaluation of a large reasoning model on the MMLU-Pro dataset. We acknowledge the comparative analysis gap with other LLMs. However, it was beyond the scope of this study. We have noted this in our Limitations section. In future studies, we plan to do a comparative analysis with multiple LLMs.

3. The paper rightly points out issues with cueing and "testwiseness" in multiple-choice questions but doesn't propose concrete mitigations. The planned future work of testing without answer choices is excellent—consider incorporating a small pilot of this now or discussing expected outcomes in more depth. Also, the limitations of using only 162 scenarios across many specialties could be made more transparent, especially regarding statistical robustness and specialty-specific insights.

Response: The manuscript addresses the problem of cueing and testwiseness in multiple-choice questions by incorporating a second study phase in which open-ended prompts were used to remove answer choices to compare against testwiseness. While the influence of testwiseness remains to be fully quantified, our findings suggest that prompt format plays a significant role in model performance, underscoring the importance of further experimentation with different assessment designs. We also note the limited sample size as a study limitation, recognizing that the use of only 162 scenarios across numerous specialties restricts statistical robustness and the depth of specialty-specific insights. Future research is warranted to improve assessment and expand the dataset size for greater generalizability.

4. The study uses a fixed prompt but does not explore or discuss the impact of prompt variations, which may influence results in open-ended tasks.

Response: We agree with the reviewer. Prompt engineering should have been undertaken early in the research process.

5. While the discussion of biases and failure modes is helpful, a more structured breakdown of error types and their frequency would improve the interpretability of findings.

Response: This is a good point; we thank the reviewer for the suggestion. We have now added a discussion of error types and frequency.

6. The discussion on reasoning steps and transparency is insightful but could be expanded to address recent concerns about the faithfulness of chain-of-thought outputs.

Response: We thank the reviewer for this comment and have now expanded the discussion on reasoning steps and chain-of-thought outputs.

Minor Comments

7. Model latency and usability: while the latency of DeepSeek is acknowledged, it's not contextualized with respect to potential clinical utility or workflow integration. A brief paragraph on practical deployment implications would strengthen the discussion.

Response: Latency of 15 to 20 seconds was experienced, and we concluded that this was inconsequential and did not any comment.

8. Citation formatting: ensure all references (especially web-based ones like Perplexity and PromptHub) are consistently formatted and maintained in the reference list.

Response: Agreed; these have been updated in the reference list.

9. Future directions could be made more actionable by suggesting benchmark expansions with real patient data or multimodal inputs.

Response: We agree; we have added a section on future implications, and we believe both of these to be viable (data and inputs).

Reviewer Q [5]

General Comments

This paper [2] seeks to evaluate the accuracy of DeepSeek R1 in correctly identifying the primary medical diagnosis in the

medical scenarios dataset portion of MMLU-Pro using an open-ended format. Some clarifications on the methods and results (especially around the roles of subject matter experts vs core team members in the publication), would be helpful in understanding how these results were derived.

Response: We have now added clarifications on the methods, and we thank the reviewer for making this suggestion to help the reader understand our results.

Minor Comments

1. Introduction: consider citing Deepseek AI's Deepseek R1 paper [6].

Response: We thank the reviewer for these additional comments. We have now included the suggested reference, as we think it will improve the context of this study.

2. Methods: please clarify who your subject matter experts were (eg, physicians, researchers) in terms of rank, specialty, and role and how they were used to grade answers (eg, selected based on specialty, 2 reviewer process, etc).

3. Methods: please indicate when the analyses were run.

Response: In the current work, since we have merged 2 papers, we have excluded the subject matter expert opinions.

4. Results: who determines whether references are related or unrelated?

Response: References were assessed by the authors of this paper; we have now clarified this.

5. Results and Discussion: it is unclear to me from reading the discussion portion of the paper as to whether we have any sense of whether DeepSeek R1 has correct reasoning for questions with correct diagnoses (eg, it may get the right diagnosis but may have incorrect reasoning). Similarly, did you determine the "correct answer" based on string matching (for example, if the answer was "septic arthritis" and the DeepSeek output stated "septic shock," would this be incorrect)?

Response: With regard to the discussion on verifying the reasoning, this is an excellent point. Unfortunately, within the scope of this current study, we were unable to assess specific reasoning steps beyond determining accuracy. String matching was not used for determining validity.

6. Discussion: consider acknowledging the sample size of questions as a limitation.

Response: We agree that the sample size was small, and we have now mentioned this as a limitation of this study.

Reviewer AA [7]

General Comments

This paper [2] *reports on an experimental study to analyze the MMLU-Pro Q&A dataset. The authors find that DeepSeek R1 had an accuracy rate of 95.1% in 162 medical scenarios after reconciliation with subject matter experts on 23 questions. The findings contribute to the growing body of knowledge on LLM applications in health care and provide insights into the strengths and limitations of DeepSeek R1 in this domain.*

Response: We appreciate the reviewer's comments, as this was our primary objective. In the current merged study, in order to use uniform formats (for both closed- and open-ended questions), we omitted the subject matter experts, removed the corresponding references and data, and recalculated the accuracy.

Major Comments

1. The results are not appropriately qualified with results on statistical significance, and/or are lacking comparisons with other language models. Even if we know how other models perform overall, it would still be good to have more details, such as a comparison of where one model is right and another is wrong. Those kinds of deep insights are lacking in this paper. All we really know is that DeepSeek performs at a level roughly equivalent to the other leading models (nothing surprising there) and that it sometimes has incomplete or inexplicable behavior. I feel the paper needs to have more results and analysis to be a good fit for this journal.

Response: Thank you for your comment. We have expanded the Results section by conducting quantitative and qualitative analysis. Conducting the analysis by using other models was beyond the scope of this study. This particular point has been listed in the Future Recommendations section. [**Editor's note:** In response to the editor's advice, the authors updated the analyses and final (accepted) version of the manuscript to include a comparison with another model (Gemini).]

2. Maybe you could add a workflow diagram/figure to better illustrate the methods?

Response: We have expanded the Methods section and and included the prompts used and an MMLU-Pro question and answer as an illustrative example. Therefore, we do not think adding a diagram would add value to the manuscript.

3. I would like Table 1 to be augmented. Perhaps you can add an example question with answer choices? Right now, it looks very trivial. The alternative is to create a simple bar graph instead of a table, but the former would be more useful.

Response: We have now removed this table from our updated, combined manuscript. We also have added an example question and associated answer choices to augment the methodological reporting of this study.

Round 2 Review

Reviewer B

My comments have been addressed.

Response: Thank you.

Reviewer Q

The paper has been revised to address the Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis–Large Language Model (TRIPOD-LLM) guidelines. Overall it appears most concerns from both reviewers have been addressed.

Response: Thank you.

References

1. Wang Z. Peer review of “The Performance of DeepSeek R1 and Gemini 3 in Complex Medical Scenarios: Comparative Study". JMIRx Med 2026;7:e96223. [doi: 10.2196/96223]
2. Bajwa M, Hoyt R, Knight D, Haider M. The performance of DeepSeek R1 and Gemini 3 in complex medical scenarios: comparative study. JMIRx Med 2026;7:e76822. [doi: 10.2196/76822]
3. Wang Z, Li H, Huang D, Kim HS, Shin CW, Rahmani AM. HealthQ: unveiling questioning capabilities of LLM chains in healthcare conversations. Smart Health (2014) 2025 Jun;36:100570. [doi: 10.1016/j.smhl.2025.100570]
4. Wornow M, Bedi S, Hernandez MAF, et al. Context clues: evaluating long context models for clinical prediction tasks on EHR data. Presented at: ICLR 2025; the Thirteenth International Conference on Learning Representations; Apr 24-28, 2025.
5. You JGT. Peer review of “The Performance of DeepSeek R1 and Gemini 3 in Complex Medical Scenarios: Comparative Study. JMIRx Med 2026;7:e96225. [doi: 10.2196/96225]
6. Guo D, Yang D, Zhang H. DeepSeek-R1: incentivizing reasoning capability in llms via reinforcement learning. arXiv. Preprint posted online on Jan 22, 2025. [doi: 10.48550/arXiv.2501.12948]
7. Kejriwal M. Peer review of “The Performance of DeepSeek R1 and Gemini 3 in Complex Medical Scenarios: Comparative Study. JMIRx Med 2026;7:e96227. [doi: 10.2196/96227]

Abbreviations

AI: artificial intelligence
LLM: large language model
MMLU-Pro: Measuring Massive Multitask Language Understanding Pro
TRIPOD-LLM: Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis–Large Language Model

Edited by A Schwartz; submitted 26.Mar.2026; this is a non–peer-reviewed article;accepted 26.Mar.2026; published 27.Apr.2026.

Please cite as:
Bajwa M, Hoyt R, Knight D, Haider M
Authors’ Response to Peer Reviews of “The Performance of DeepSeek R1 and Gemini 3 in Complex Medical Scenarios: Comparative Study”
JMIRx Med 2026;7:e96220
URL: https://xmed.jmir.org/2026/1/e96220
doi:10.2196/96220

Author's Response to Peer Reviews of "Interpreting the Estimand Framework From a Causal Inference Perspective"

Jinghong Zeng[1,2], MSc

[1]Department of Statistics, University of Auckland, 38 Princes Street, Auckland, New Zealand

[2]Department of Statistics and Programming, Jiangsu Hengrui Medicine (China), Guangzhou, Guangdong, China

Corresponding Author:
Jinghong Zeng, MSc
Department of Statistics, University of Auckland, 38 Princes Street, Auckland, New Zealand

Related Articles:

Companion article: http://preprints.jmir.org/preprint/88813

Companion article: https://med.jmirx.org/2026/1/e98122

Companion article: https://med.jmirx.org/2026/1/e98125

Companion article: https://med.jmirx.org/2026/1/e98126

Companion article: https://med.jmirx.org/2026/1/e88813

(JMIRx Med 2026;7:e98121) doi:10.2196/98121

KEYWORDS

causal inference; clinical trial; estimand; intercurrent event; treatment effect

This is the author's response to peer-review reports for "Interpreting the Estimand Framework From a Causal Inference Perspective [1]."

Round 1 Review

Reviewer G [2]

Major Comments

1. *The professional society "ICH" is never spelled out or introduced. Additional context is needed regarding the role of the ICH, its influence on regulatory science, and why its guidelines are particularly important for clinical trial design and analysis.*

Response: I've added an introduction to the International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use (ICH) to the first paragraph of the Introduction section.

2. *The Efficacy Guideline E9 was published in 2019. The authors should clarify what impact this guideline has had on the pharmaceutical industry since its release.*

Response: I've added an introduction to E9 and E9 (R1) to the first paragraph of the Introduction section.

2. *Moreover, it is unclear why a causal interpretation of this guideline is timely and important in 2025, several years after its publication.*

Response: I added some justifications for this aim in the second-, third-, and fourth-to-last paragraphs. The estimand is different from a causal treatment effect, so a causal interpretation could be helpful.

3. *None of the proposed strategies address noncompliance, such as cases where treatment is not received despite assignment or is received without assignment (eg, X(R=1)=0 or X(R=0)=1). Noncompliance is a central issue in causal inference and should be explicitly discussed. If noncompliance is assumed to be irrelevant, then the introduction of the notation R appears redundant and should be justified or removed.*

Response: Yes, noncompliance is an important issue in causal inference. Since causal inference is a broad area, this article is not intended as a review of causal inference. Only unmeasured confounding is discussed as an example. Noncompliance is not discussed further, but is mentioned in the Introduction section. The notation R is needed for discussion of unmeasured confounding and the intention-to-treat (ITT) principle.

4. *The strategies are presented at a very high level. Although the 4 cases illustrated in Figure 2 provide some intuition regarding the appropriateness of each strategy, the Viewpoint would be substantially strengthened by grounding the discussion in real clinical trial examples. Demonstrating how each strategy has been applied in practice would greatly improve clarity and impact.*

Response: I've added real clinical trial examples and explained them.

5. *The scope and framing of the Viewpoint appear better suited for a pharmaceutical science or regulatory-focused journal rather than a JMIR-based journal. The authors should better justify the relevance of this work to the JMIR readership or reconsider the target venue.*

Response: I think the paper suits the broad audience of this journal because interpretation from a causal inference perspective works not only for clinical trials in the pharmaceutical industry, but also for clinical trials in academic and observational studies. I've added a justification in the Introduction section.

Minor Comments

1. *Section 2 begins with the statement: "A causal inference framework is based on the potential outcome framework." This is inaccurate, as causal inference can also be grounded in other frameworks, such as structural causal models.*

Response: Sorry for the inaccuracy. I only meant to say that the potential outcome framework is discussed. I've modified this sentence and added a clarification to the Introduction section.

2. *In the abstract, the sentence "This article aims to interpret the estimand framework through its underlying theories, the causal inference framework based on potential outcomes" should replace the comma with "and" for grammatical correctness.*

Response: I've revised the abstract and deleted this sentence.

3. *On page 2, second line: "Generally, Treaments are…" — the "T" in "Treatments" should not be capitalized.*

Response: I've removed the capitalization of this letter.

4. *In section 3.2 (page 6): the sentence "Through the hypothetical strategy, we make the second…" is ambiguous and should be rewritten for clarity.*

Response: I've rewritten the discussion about intercurrent event strategies. I've deleted this sentence.

Reviewer H [3]

Major Comments

1. *The introduction may give the impression that these strategies are newly proposed by the author, whereas they are in fact defined in ICH E9 (R1). The manuscript would benefit from clearer attribution to, and positioning relative to, the ICH E9(R1) estimand framework.*

Response: I have clearly attributed the estimand framework to ICH E9 (R1) in the Introduction section.

2. *The important concept of intercurrent events is not clearly defined. The definition provided in the manuscript, "Intercurrent events are events that happen after treatment initiation and affect the definition of a treatment effect" (page 2) is vague and potentially misleading. It misses the key idea that intercurrent events are posttreatment events that interfere with the interpretation or existence of the outcome relative to the treatment of interest, rather than merely events that affect treatment effects.*

Response: I've improved the definition of intercurrent events according to the original definition in ICH E9 (R1) in the Introduction section.

3. *In section 2, it is incorrect to state that "Ri, Xi and Yi are potential outcomes." Only Xi (·) and Yi (·) are potential outcomes. The randomization indicator Ri not a potential outcome; it is a realized random variable determined by the design.*

Response: I have not made changes. I disagree with this comment because the randomization scheme R here is also a potential variable. When we imagine a participant being randomized to a treatment arm or a control arm, we are imagining two different randomization outcomes. A participant cannot be assigned to two arms at the same time, so the two randomization outcomes are potential: they have not yet been realized when we consider potential outcomes of X and Y.

4. *At the beginning of section 2, the authors assume "an ideal two-arm randomized controlled clinical trial, with full compliance to treatment and no intercurrent events." In such a setting, confounders do not affect treatment assignment. However, the manuscript later defines "some confounders C that affect both X and Y," which contradicts the assumption of randomization.*

Response: Yes, I agree. To make the description more accurate, I have deleted the word "ideal." I have also deleted "and no intercurrent events."

5. *Average treatment effect (ATE) is defined as ATE = E(Y (X(R = 1) = 1) | C) − E(Y (X(R = 0) = 0) | C). However, this is a conditional ATE rather than the marginal ATE, since Y (X(R = 1) = 1) and Y (X(R = 0) = 0) are potential outcomes. The author should define the ATE marginally and then mention conditioning on C for adjustment.*

Response: Initially, I tried to write the methods in a simple mode to help readers understand, but this reduced some theoretical accuracy. I have fully updated the description of the causal inference framework with more rigorous research findings from my recently published paper. The ATE now is marginal for the whole article.

6. *Page 3: the phrase "the difference (D) between the average treatment effect from participants who take the experimental treatment..." is incorrect. The quantity described is the difference in average observed outcomes, not an average treatment effect.*

Response: I've updated the description of the causal inference framework. I deleted D.

7. *Across strategies, the author repeatedly claims that the estimand formula is "still" the same, which is misleading. The symbolic form may look similar, but the estimand is not the same. In treatment policy, X is redefined; in composite and while-on-treatment strategies, Y is redefined; in principal stratification, the target population changes. This undermines*

the central E9 (R1) message that different strategies define different estimands.

Response: Sorry for the lack of clarity. I did not intend to undermine the estimand framework. In fact, I think highly of it. I have rewritten all descriptions of intercurrent event strategies.

8. *The proposed "model adjustment strategy" does not correspond to an estimand strategy as defined in ICH E9 (R1), but rather to a particular modeling or estimation approach. Moreover, in case 1 of Figure 2, concomitant therapies occur after treatment initiation, which is inconsistent with the causal diagram in Figure 3. In this setting, M may act as a mediator rather than a confounder. Treating postrandomization intercurrent events as confounders requires careful causal justification and may induce bias; this issue is not discussed in the manuscript.*

Response: Sorry for the oversimplification. After consideration, concomitant therapies after treatment initiation may indeed act as a mediator and also as a confounder for subsequent treatment. Indeed, this case can be complicated. After revision, I think the model adjustment strategy does not fit the article focus well, so I have deleted this section.

Minor Comments

9. *Please spell out the abbreviation "ICH" at its first occurrence.*

Response: I've spelled out ICH.

10. *Some sentences are confusing and would benefit from revision. For example, in the second paragraph on page 2: "Intercurrent events are frequent in practice but conceptually novel. E9(R1) listed many examples for intercurrent events, such as use of concomitant therapies, treatment switching and death before endpoint measurement." Intercurrent events are not really new conceptually; rather, they were newly formalized or explicitly emphasized in E9 (R1). In "examples for intercurrent events," the preposition should be "of," not "for." As a second example, "This individual treatment effect controls confounders on the endpoint within the same participant and means how the endpoint would change when only the treatment condition changes" on page 3: the ITE does not "control confounders"; it is defined counterfactually for the same individual.*

Response: I have revised "Intercurrent events are frequent in practice but conceptually novel. E9(R1) listed many examples for intercurrent events, such as use of concomitant therapies, treatment switching and death before endpoint measurement." The new version is as follows: "Intercurrent events are very common in practice, including use of concomitant therapies, treatment switch and death before endpoint measurement." Also, I have revised "This individual treatment effect controls confounders on the endpoint within the same participant and means how the endpoint would change when only the treatment condition changes" as follows: "The ITE means how the endpoint would change when only the treatment condition changes for this participant," where ITE has been spelled out.

11. *A right parenthesis is missing in the first ATE formula on page 3.*

Response: Now the formula is complete.

12. *Equation 2.1 is missing the observed randomization indicator R^o in the first line.*

Response: I have deleted this equation, since I have updated the causal inference framework description.

13. *The exclusion restriction assumption for instrumental variables should be stated more clearly.*

Response: Many important assumptions are not mentioned in this article, as they are very technical. I have provided a reference to a complete list of assumptions instead.

Reviewer P [4]

Major Comments

1. *The manuscript repeatedly states that it "interprets" the ICH E9 framework, but in practice, it mostly rephrases ICH E9 concepts using potential outcomes notation. Readers would more likely expect to see discussions on limitations, ambiguities, or contested aspects.*

Response: I have added more discussion of real clinical trial examples. I have also discussed differences between estimands and causal treatment effects.

2. *While pedagogical simplicity may be intentional, several aspects risk being misleading if read uncritically. For example, conditioning on posttreatment variables (section 3.6) is introduced without adequate warning about collider bias or causal ordering issues, and the discussion of principal stratification glosses over identification challenges, relying on brief mentions of Bayesian methods without clarifying assumptions. These are not fatal flaws, but the author should be more explicit about what is heuristic versus formally justified.*

Response: Yes, the discussion of conditioning on posttreatment variables requires more detail. Sorry for the oversimplification. However, this topic became less relevant to the revised paper, so I have now deleted this section. The article now focuses on a comparison between estimands and causal treatment effects, so the discussion of principal stratification now primarily serves this focus, and modeling approaches are not mentioned anymore.

References

1. Zeng J. Interpreting the estimand framework from a causal inference perspective. JMIRx Med 2026;7:e88813. [doi: 10.2196/88813]
2. Zhang L. Peer review of "Interpreting the Estimand Framework From a Causal Inference Perspective". JMIRx Med 2026;7:e98122. [doi: 10.2196/98122]

3. Zhang Q. Peer review of "Interpreting the Estimand Framework From a Causal Inference Perspective. JMIRx Med 2026;7:e98125. [doi: 10.2196/98125]
4. Wu H. Peer review of “Interpreting the Estimand Framework From a Causal Inference Perspective”. JMIRx Med 2026;7:e98126. [doi: 10.2196/98126]

Abbreviations

ATE: average treatment effect
ICH: International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use
ITT: intention-to-treat

Edited by A Schwartz; submitted 13.Apr.2026; this is a non–peer-reviewed article;accepted 13.Apr.2026; published 22.May.2026.

Please cite as:
Zeng J
Author’s Response to Peer Reviews of “Interpreting the Estimand Framework From a Causal Inference Perspective”
JMIRx Med 2026;7:e98121
URL: https://xmed.jmir.org/2026/1/e98121
doi:10.2196/98121

Use of a Specialist Telephone Consultation Line for Long COVID in Primary Care in British Columbia: Retrospective Descriptive Quality Improvement Study

Saniya Kaushal[1,2], BCh, BAO (Hons), MB; Jastinder Bhandal[3], BKin; Peter Birks[4], BMSc, MD, MHA; Jesse Greiner[4], BSc, MSc, MD, MBA; Adeera Levin[4], MD; Michelle Malbeuf[1], BSCN, MHA; Zachary Schwartz[4], BSc, MD

[1]Providence Health Care Research Institute and Provincial Health Services Authority, 1081 Burrard Street, Vancouver, BC, Canada

[2]Postgraduate Medical Education – Internal Medicine, Toronto Metropolitan University, Toronto, ON, Canada

[3]School of Medicine, University of Limerick, Limerick, Ireland

[4]Faculty of Medicine, The University of British Columbia, Vancouver, BC, Canada

Corresponding Author:
Saniya Kaushal, BCh, BAO (Hons), MB
Providence Health Care Research Institute and Provincial Health Services Authority, 1081 Burrard Street, Vancouver, BC, Canada

Related Articles:

Companion article: https://www.medrxiv.org/content/10.1101/2024.01.28.24301875

Companion article: https://med.jmirx.org/2026/1/e89735

Companion article: https://med.jmirx.org/2026/1/e90935

Companion article: https://med.jmirx.org/2026/1/e89710

Abstract

Background: Long COVID (post-COVID-19 condition) continues to challenge primary care. To support family physicians in British Columbia, the general internal medicine (GIM) COVID-19 Rapid Access to Consultative Expertise (RACE) line was launched in August 2020 to provide real-time specialist advice.

Objective: This quality improvement study aimed to evaluate the implementation and utilization of the GIM-COVID-19 Long-Term Sequelae RACE line in British Columbia. Specifically, it sought to characterize the demographics of patients involved in RACE consultations, identify the most common themes and clinical queries presented by primary care providers, and assess how usage patterns evolved over time during the COVID-19 pandemic.

Methods: We conducted a retrospective descriptive analysis of 149 RACE line call summaries between August 2020 and June 2021. Six calls were excluded due to insufficient information, such as incomplete documentation or absence of a clear COVID-19–related question. Because the original extraction notes are no longer available, further details about these calls cannot be provided, leaving 143 eligible calls. Data extracted included patient age, sex, geographical location, symptom type, and timing of symptom onset post–COVID-19 infection. Calls were categorized by symptom duration (acute: <2 wk, subacute: 2 - 12 wk, chronic: >12 wk), thematic content (respiratory, fatigue, neurological, etc), and query type (symptom management, return-to-work, vaccination, etc). Data were coded independently by two reviewers using a standardized spreadsheet and predefined codebook. Discrepancies were resolved through discussion. Descriptive statistics summarized the findings.

Results: Many calls involved female patients (91/143, 64%), with the most common age group being 40 - 49 years (32/113, 28%). Most calls came from Greater Vancouver (35/83, 42%) and the Fraser Valley (29/83, 35%). Subacute symptoms (52/149, 35%) and vaccination-related concerns (29/149, 19%) were the most common inquiry types. Symptom-related inquiries accounted for 92 of 143 calls (64%), with 253 symptoms documented overall. Respiratory symptoms were most common (100/253, 40%), especially shortness of breath (35 calls), cough (26), and fatigue (23). Call volumes peaked from January to June 2021, coinciding with the provincial vaccine rollout.

Conclusions: The GIM-COVID-19 Long-Term Sequelae RACE line served as a critical early support system for primary care providers as the long COVID landscape evolved. This quality improvement study emphasizes the value of rapid access and specialist-informed consultation tools during emerging public health challenges. The trends ascertained may inform future health

system responses, particularly when designing more scalable, interdisciplinary models to support primary care in managing complex chronic conditions.

(JMIRx Med 2026;7:e57021) doi:10.2196/57021

KEYWORDS

internal medicine; long COVID; COVID-19; SARS-CoV-2; GP; general practice; general practitioner; consult; respiratory; infectious; respiration; primary care; telephone; telehealth

Introduction

Background

As COVID-19 transitions from a pandemic to endemic, its long-term effects, commonly referred to as long COVID, continue to place a substantial burden on both patients and health care systems. Real-time access to guidance for physicians has been offered through Rapid Access to Consultative Expertise (RACE) in British Columbia for over 12 years [1]. In August 2020, this service was expanded to include a long COVID RACE line, designed to support primary care physicians managing patients with persistent symptoms. Patients experiencing residual symptoms, such as shortness of breath, cough, and fatigue, for months postinfection are being diagnosed with postacute sequelae of SARS-CoV-2 (PASC) [2]. The World Health Organization defines post-COVID-19 condition as symptoms lasting at least two months and occurring usually three months from the onset of COVID-19, which cannot be explained by an alternative diagnosis [3]. In this study, data regarding patients with PASC collected through the general internal medicine (GIM)–COVID-19 Long-Term Sequelae division of the provincial RACE line were tabulated to identify trends in the long-term progression of COVID-19 symptoms. Findings of this study will offer information to improve the provincial patient care and long-term support for future patients with COVID-19, while offering insights on the usage of the GIM-COVID-19 Long-Term Sequelae RACE line.

As of June 27, 2022, the COVID-19 pandemic had affected over 544 million people worldwide [4]. The virus responsible for COVID-19 is SARS-CoV-2 [2]. Between its detection in December 2019 and June 27, 2022, the COVID-19 virus had infected 3.94 million Canadians [4]. British Columbia accounted for 373,974 of these cases [5]. Although over 369,000 British Columbians recovered from their acute COVID-19 illness, many are still experiencing residual symptoms for months or longer [5]. These patients have been deemed to have PASC [6]. This syndrome has many other names too, including long COVID, with some patients calling themselves long haulers; it is heterogeneous in presentation [2,6]. A 2022 *BMJ Open* meta-analysis approximated global PASC occurrence at 54% in hospitalized individuals and 34% in nonhospitalized patients, with an overall pooled occurrence of 43% [7]. In one study, nearly one-third of people who recovered from COVID-19 said they were still dealing with lingering symptoms that affected their day-to-day quality of life. Furthermore, recent US data suggests that about 1 in 7 adults have experienced symptoms of long COVID [8]. Common symptoms of long COVID include fatigue, dyspnea, chest pain, palpitations, cognitive dysfunction (“brain fog”), and anxiety, which may fluctuate or persist for months after acute infection [9]. Long COVID often flies under the radar in primary care. For example, in a large group of patients in England who had confirmed COVID-19, only about 1.8% were officially recorded as having long COVID, suggesting that many cases might be going unrecognized in everyday practice [10].

The underlying pathophysiology of long COVID is still being studied. However, contributions from immune dysregulation, viral persistence, microvascular dysfunction, and autonomic nervous system imbalance have been noted in several studies [11]. As of 2024, the National Institutes of Health RECOVER Initiative in the United States is conducting 8 clinical trials evaluating 13 potential interventions across 5 key symptom areas, with studies launched at over 50 sites nationwide to investigate treatments for long COVID [12,13]. Despite the increasing research, primary care providers often lack confidence in managing long COVID. A 2023 survey of 53 general practitioners (GPs) in Ireland found that only 8% felt confident in diagnosing long Covid and 81% were not confident in managing it, with 70% unaware of referral indications and 93% reporting educational gaps [14]. These findings highlight the pressing need for well-defined referral pathways and timely specialist involvement to support GPs in managing this complex condition effectively.

In British Columbia, Canada, the Post-COVID-19 Interdisciplinary Clinical Care Network (PC-ICCN) was developed to support the best outcomes for patients recovering from symptoms following COVID-19 infection through research, education, and clinical care [15]. One of the clinical resources within the PC-ICCN included the establishment of the GIM-COVID-19 Long-Term Sequelae division of the provincial RACE line [15]. This resource provides immediate (<2 h) specialist advice to GPs caring for patients with long-term sequelae of COVID-19 infection. The provincial RACE line has existed in British Columbia since 2010 and provides access to immediate specialist advice/consultation across the province [1].

The GIM-COVID-19 Long-Term Sequelae RACE line is answered by a dedicated group of GIM specialists with an interest in and experience with acute and chronic COVID-19 [1]. The guidance provided by the GIM physicians includes diagnostic investigations, management, and navigation of these complex patients. Since its inception, the RACE line has aimed to bridge gaps in access to timely specialist input as an approach intended to reduce unnecessary referrals and support primary care providers in managing complex conditions [16]. In Ontario, a special eConsult service for long COVID helped family doctors get quick advice from specialists, which meant patients

got the help they needed faster, often without having to see a specialist in person [17].

Similar post-COVID advice or consultation pathways have been established in countries such as the United Kingdom (NHS long COVID clinics), the United States (National Institutes of Health RECOVER Initiative), and Australia (long COVID management guidelines for GPs), reflecting global recognition of the condition's complexity and the need for specialist support [10,18,19].

This report presents an analysis of the types and frequencies of calls made to the GIM-COVID-19 Long-Term Sequelae RACE line. This analysis enables the identification of trends in patient presentations, primary care practitioner concerns, and related questions. This data informs on the development of education, tools, and care plans, which improves the quality of care and long-term support for patients with COVID-19 and their health care providers.

Objectives

The aim of this study was to evaluate the usage patterns, themes of inquiry, and demographic data associated with the GIM-COVID-19 Long-Term Sequelae RACE line in British Columbia. By analyzing call content and frequency, we sought to identify knowledge gaps among primary care providers and inform future improvements in post–COVID-19 care resources and communication strategies.

Methods

Overview

This is a quality improvement study evaluating the GIM-COVID-19 Long-Term Sequelae RACE line data from its launch in August 2020 to June 2021. These data are comprised of the documented exchanges between primary care practitioners (PCPs) and GIM specialists. The study received quality improvement approval through Providence Health Care and was exempt from formal research ethics board review. No patient-identifying data were accessed, and all analysis was conducted on anonymized call notes.

Data Source and Call Selection

In total, 149 RACE line call medical notes were systematically reviewed to extract data regarding the variables of interest: patient demographics (age, sex, region) and types of queries related to COVID-19 (acute symptoms, subacute symptoms, chronic symptoms, vaccination inquiries, miscellaneous questions). The data from these calls were tabulated for analysis. Call notes were reviewed manually using a standardized spreadsheet for data extraction. Each variable was assigned a predefined codebook category, and disagreements were resolved through discussion and consensus. Six calls were excluded from this study because they were too vague to draw conclusive findings. Reasons for exclusion included incomplete documentation, lack of a clear COVID-19–related question, or insufficient clinical information to categorize the inquiry. Excluded calls were logged and reviewed to ensure consistent application of exclusion criteria. The remaining 143 calls were used to observe trends in age, sex, geographical location, types of queries, timing, and symptoms of patients of the GIM-COVID-19 Long-Term Sequelae RACE line between August 2020 and June 2021. All calls categorized under the GIM-COVID-19 Long-Term Sequelae RACE line service during this period were included; calls unrelated to post-COVID symptoms were excluded.

Query Classification

For RACE calls regarding patient symptoms, we examined the reported symptoms according to the time post–COVID-19 infection. This was predetermined as 0 - 2 weeks following diagnosis to represent acute COVID-19 symptoms, 2 - 12 weeks following diagnosis to represent subacute COVID-19 symptoms, and >12 weeks following COVID-19 diagnosis to represent chronic COVID-19 symptoms. The relative frequencies of types of symptoms reported to the GIM-COVID-19 Long-Term Sequelae RACE line were analyzed and compared.

Temporal Analysis of Calls

We also analyzed the reasons for calls to the RACE line by time period within the pandemic. First, we described the type of calls received during the different COVID-19 "waves" as occurred in British Columbia, including August to December 2020, January 2021 to March 2021, and April 2021 to June 2021. Second, we assessed the reasons for the call before and after the availability of COVID-19 vaccines. We hypothesized that the reasons for calls would vary depending on the time period during the pandemic. In addition, calls were ranked by the phase of the pandemic to explore trends in the volume and nature of queries, including vaccine-related concerns and the timing of chronic symptom presentations. Descriptive statistics (frequencies and proportions) were used to summarize query types by time period. The data were analyzed using Microsoft Excel (Microsoft Corp), which was then used to organize, summarize, and identify patterns in query types, symptom categories, and temporal trends.

Symptom Categorization and Coding

Each call was examined to determine whether it included symptom-related content, which was then grouped by organ system (eg, respiratory, neurological, or gastrointestinal). In cases where more than one symptom was mentioned, all relevant details were recorded to capture the full scope of the patient's concerns. When symptoms overlapped or were unclear, the team discussed them collectively before assigning categories to maintain consistency across the dataset.

Calls were coded based on symptom duration, query type, and organ system using a predefined framework (Table 1). The initial framework was adapted from published sources and refined after the pilot coding of 10 calls. Two reviewers independently applied this framework to the complete dataset, meeting regularly to compare interpretations and update the categories as new themes appeared. Any differences in coding were resolved through discussion until consensus was reached.

Table . Coding framework used to classify symptom duration, query type, and symptom system for RACE line calls.

Variable	Categories	Description/examples
Symptom duration	Acute (<2 wk), subacute (2 - 12 wk), chronic (>12 wk)	Based on time since acute COVID-19 infection
Query type	Symptom management, vaccination, return-to-work/school, diagnosis clarification, medication advice, other	Categorized according to the primary purpose of the call
Symptom system	Respiratory, fatigue, neurological, cardiovascular, mental health, gastrointestinal, multisystem, other	Grouped by system affected, based on physician documentation

Although formal interrater reliability statistics (such as κ values) were not calculated, the team held multiple calibration meetings to ensure categories were applied consistently. All coded data were entered into a structured Microsoft Excel sheet for organization and analysis. Quantitative findings were summarized descriptively, while qualitative insights were drawn inductively from the free-text notes accompanying each call.

Thematic Analysis

A qualitative content analysis was carried out to explore recurring themes in the clinical questions and patient symptom narratives. Rather than relying on a preset framework, themes were drawn directly and inductively from the anonymized call notes. This approach helped the research team capture how primary care providers' clinical concerns and informational needs evolved over time while managing patients with post–COVID-19 conditions during the study period.

Ethical Considerations

This project was conducted as a quality improvement initiative and did not involve direct interaction with patients or the collection of personal patient information that would allow identification. Thus, it did not require formal review by a research ethics board. This project was reviewed and approved as a quality improvement initiative by Providence Health Care and was deemed exempt from formal institutional research ethics board review.

All the data analyzed were anonymized call summaries obtained from the RACE line, which documents virtual consultations between PCPs and GIM specialists. No identifiable or sensitive personal health information was accessed, and there were no links to patient charts or follow-up data.

Individual informed consent was not required since this was a secondary analysis of anonymized and nonidentifiable data. No participants were directly involved or contacted for the purposes of this study. The original data source did not include participant-level identifiers or contact information. No compensation was offered to anyone, as this study did not include human subjects or involve any form of participant recruitment. No images or materials in this manuscript include identifiable individuals.

Results

Patient Demographics

Table 2 outlines the characteristics of patients whose primary care providers contacted the RACE line between August 2020 and June 2021. The majority of calls concerned patients in middle age, most commonly those between 40 and 49 years (28%), followed by individuals aged 50 - 59 years (23%) and 30 - 39 years (16%). Women represented nearly two-thirds of all patients (64%), while men accounted for just over one-third (36%).

Geographically, calls were concentrated in the more urban and suburban regions of British Columbia, particularly Greater Vancouver (35/83, 42%) and the Fraser Valley (29/83, 35%)—together making up more than three-quarters of all consultations. Smaller proportions of calls came from Vancouver Island (8/83, 10%), with the remainder originating from rural and remote regions, including Northern British Columbia (4/83, 5%) and the Interior (3/83, 4%), as well as from out of province (Yukon: 3/83, 4%; Alberta: 1/83, 1%).

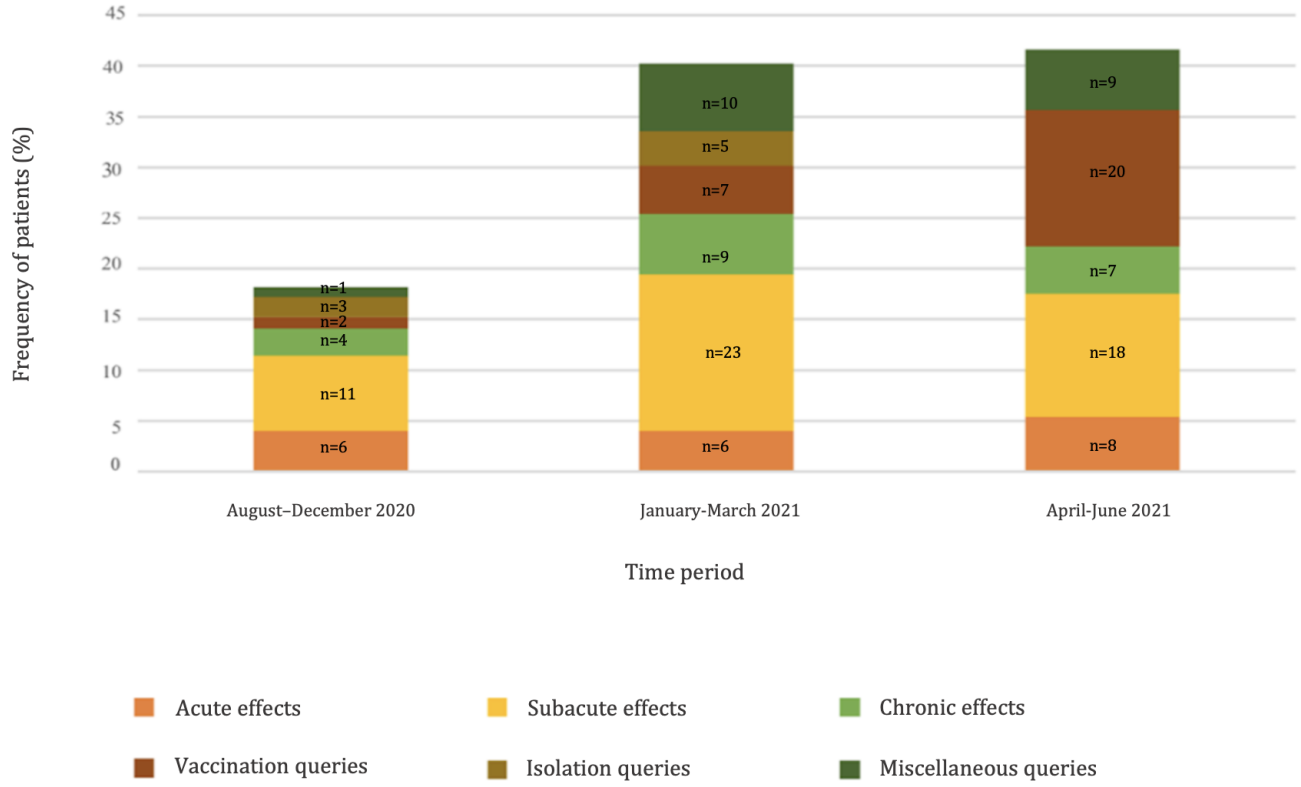
Table . Demographics of COVID-GIM-Post-Infection Care RACE line patients. Percentages may not total 100 due to rounding.

Variable	Category	Patients, n (%)
Age group, years (n=143)		
	<20	3 (2)
	20 - 29	8 (7)
	30 - 39	18 (16)
	40 - 49	32 (28)
	50 - 59	26 (23)
	60 - 69	9 (7)
	70 - 79	13 (12)
	≥80	5 (4)
Sex (n=143)		
	Male	52 (36)
	Female	91 (64)
Geographic region (n=83)		
	Greater Vancouver	35 (42)
	Fraser Valley	29 (35)
	Vancouver Island	8 (10)
	Northern British Columbia	4 (5)
	Interior British Columbia	3 (4)
	Yukon	3 (4)
	Alberta	1 (1)

Types of Queries Received

The types of COVID-19–related queries that were received by the COVID-GIM-Post-Infection Care RACE line are summarized by time period. The data demonstrate that subacute (2 - 12 wk following diagnosis) symptoms were cited in 52 of 149 calls (35%) and vaccination queries were cited in 29 calls (19%), making both the most common RACE line call matters. A visual breakdown of these query types by time period is provided in Figure 1.

Figure 1. Progression of types of COVID-19–related queries to the COVID-GIM-Post-Infection Care RACE line by time period. RACE: Rapid Access to Consultative Expertise.



Timing of RACE Line Queries

As shown in Figure 1, there was a much larger frequency of calls from January to March 2021 (n=60, 40%) and April to June 2021 (n=62, 42%) compared to the earlier 27 (18%) calls from August to December 2020. There was a much larger number of queries following the introduction of the COVID-19 vaccine (n=122, 82%) compared to the time period prior (n=27, 18%). This increase in queries temporally coincided with the start of the vaccine rollout in early 2021.

Symptoms Reported by System and Time Frame

Symptoms reported to the GIM-COVID-19 Long-Term Sequelae RACE line between August 2020 and June 2021 were compiled. Data from the calls that reported symptoms present 2 - 12 weeks following COVID-19 diagnosis (subacute) or >12 weeks following COVID-19 diagnosis (chronic) are summarized visually in Table 3.

Table . Symptoms reported to the COVID-GIM-Post-Infection Care Rapid Access to Consultative Expertise line by system and category.

System/category	Values, n (%)
Subacute symptoms (2 - 12 wk following diagnosis)	
Respiratory	71 (28.1)
Malaise/fatigue	35 (13.8)
Neurological	17 (6.9)
Cardiac	16 (6.4)
Musculoskeletal	10 (3.9)
Gastrointestinal	6 (2.5)
Neuropsychological	6 (2.5)
Abnormal laboratory/diagnostic findings	0 (0)
Genitourinary	4 (1.5)
Autoimmune	0 (0)
Dermatological	0 (0)
Chronic symptoms (≥12 wk following COVID-19 diagnosis)	
Respiratory	29 (11.3)
Malaise/fatigue	11 (4.4)
Neurological	9 (3.4)
Cardiac	9 (3.4)
Musculoskeletal	9 (3.4)
Gastrointestinal	5 (2)
Neuropsychological	4 (1.5)
Abnormal laboratory/diagnostic findings	5 (2)
Genitourinary	0 (0)
Autoimmune	3 (1)
Dermatological	3 (1)
Combined frequency of reported symptoms following COVID-19 infection (n=253)	
Respiratory	100 (39.5)
Malaise/fatigue	46 (18.2)
Neurological	26 (10.3)
Cardiac	25 (9.9)
Musculoskeletal	19 (7.5)
Gastrointestinal	11 (4.3)
Neuropsychological	10 (4)
Abnormal laboratory/diagnostic findings	5 (2)
Genitourinary	4 (1.6)
Autoimmune	3 (1.2)
Dermatological	3 (1.2)

Questions about specific symptoms varied by time postinfection, but the most frequent symptoms across time periods included shortness of breath, cough, and fatigue. Overall, 92 (62%) calls focused on symptoms; within these calls, there were 253 total symptoms reported. Shortness of breath was the most reported symptom, being identified in 35 calls (38%). Cough, fatigue, and fever were the other highest reported symptoms recorded among GIM-COVID-19 Long-Term Sequelae RACE line patients, accounting for 26 calls (28%), 23 calls (25%), and 22 calls (24%), respectively. The most common symptoms reported 12 weeks postinfection were shortness of breath, chest pain, and fatigue.

Table 3 organizes the symptoms by body system, allowing for recognition of symptom clusters and trends. Our findings demonstrated that respiratory, malaise/fatigue, and neurological symptoms were the most common categories of postinfection symptoms resulting in calls to the RACE line. Respiratory symptoms made up 40% (100/253) of reported symptoms. Respiratory symptoms reported included cold, coryza, cough, hypoxemia, lung infiltrate, nasal congestion, rhinorrhea, shortness of breath, sore throat, sputum, and wheezing. Shortness of breath, cough, and hypoxemia were the most persistent respiratory symptoms listed. Neurological symptoms described by PCPs regarding their patients included dizziness, light-headedness, numbness, vertigo, and paresthesia. The wide range of symptoms reported reinforces the multisystemic nature of long COVID syndromes and the complexity of clinical management in primary care settings.

Discussion

RACE Line Utilization and Demographics

This report provides information on the use of the GIM-COVID-19 Long-Term Sequelae RACE line, specifically between August 2020 and June 2021. Our data suggest that the RACE line has been a utilized resource for PCPs, especially in the Greater Vancouver and Fraser Health regions. The data obtained indicate that the frequency of calls to the RACE line has increased throughout the pandemic. Although this RACE line is accessible across the province, there was minimal uptake outside the two aforementioned regions, as seen in Table 2. This may reflect "burden" and populations affected or an underutilization by more sparsely populated regions. Further analysis of unmet need versus not needed is required so that we can ascertain if different strategies for awareness of the RACE line outside of highly populated areas is required. This study highlights the need to understand mechanisms by which GPs learn about new RACE lines and this resource in particular. These findings highlight the opportunity to use centralized consultation services like the RACE line more strategically during public health crises. As novel health conditions emerge, having specialist-access infrastructure already in place can ensure faster, provincewide dissemination of clinical support.

Observable trends in the data indicated that the most common age group of RACE line patients was 40 - 49 years old, with 40 of 143 patients (28%) in this bracket. Females made up 91 of 143 patients (64%). Interestingly, these demographics are also representative of those mostly likely to be referred and seen in post-COVID recovery clinics across the province. Thus, information from clinicians to PCPs is based on relevant experience in a similar population. Future planning should include population-based targeting strategies and outreach initiatives tailored to groups underrepresented in utilization patterns, including rural and Indigenous communities.

Themes in Queries

RACE line subject matter involved acute, subacute, and chronic effects experienced by the patients. Respiratory, malaise/fatigue, and neurological symptoms were the most common categories of postinfection symptoms reported to the RACE line. This data can be used to identify gaps in PCP knowledge in the diagnosis and management of persistent symptoms following COVID-19 infection and has informed the development of educational resources for health care practitioners. Adopting these care plans would significantly improve the quality and aid provided by future RACE line calls in this division. Therefore, increasing the accessibility of resources on managing these identified symptoms should be a focus moving forward. This is an initiative that is currently being implemented by the provincial post-COVID recovery clinic websites. Other topics of interest include that RACE line calls also frequently centered around vaccinations, isolation periods, the impact of preexisting health conditions on COVID-19 manifestation, and antiviral treatments. This is despite regular bulletins and updates to medical doctors from provincial health bodies, infectious disease specialists, and public health officers. This highlights key areas of uncertainty and opportunities for clearer messaging to health care practitioners. Our findings align with global surveys of GPs, such as those referenced in the Introduction, that identified a lack of confidence in managing long COVID, particularly around referral pathways and symptom assessment. The RACE line model may serve as a blueprint to help bridge these knowledge gaps and offer just-in-time support when new syndromes with unclear management approaches arise.

Our findings are consistent with trends observed internationally. In British Columbia, the RACE line was most often used for queries related to respiratory, fatigue, and neurological symptoms, as well as questions about vaccination and isolation. These areas of uncertainty are not unique to the province. Similar issues have been described elsewhere, prompting the creation of virtual and rapid-access models for post–COVID-19 care.

In the United Kingdom, Leeds launched one of the earliest integrated long COVID programs that combined a specialist multidisciplinary team, community rehabilitation services, and self-management resources, reflecting NHS England's broader plan to expand post-COVID assessment clinics [20]. By late 2021, the NHS had established nearly 90 dedicated long COVID clinics across England, providing comprehensive multidisciplinary assessment and rehabilitation [21].

In the United States, the Johns Hopkins Post-Acute COVID-19 Team used telemedicine to coordinate care across multiple specialties without relying on a centralized physical clinic. This approach highlighted how virtual models can effectively manage complex cases and maintain accessibility [22].

In Australia, a Melbourne-based telehealth long COVID service has supported more than 500 patients nationwide, including those in rural and pediatric populations. Its collaboration with local primary care providers helped ensure equitable access despite geographical barriers [23].

Across Europe, similar models have been implemented. A national survey of 124 post-COVID clinics in Italy found that 93.5% maintained direct communication pathways with GPs, and nearly one-quarter incorporated telemedicine into their standard care processes [24].

Together, these initiatives reflect a shared global recognition of the need for coordinated, multidisciplinary approaches to

long COVID care. Although the RACE line differs as a physician-to-physician consultation model, it aligns with these international strategies by promoting timely access to specialist input and integration within primary care networks.

Study Strengths and Limitations

This study examines routinely standardized and collected information generated from RACE calls over an 11-month period of time, where a small number of dedicated individuals were answering calls. To our knowledge, this is the only provincial post-COVID RACE line set up in Canada. Of the 149 RACE line calls made to the GIM-COVID-19 Long-Term Sequelae RACE line during the time period being investigated**,** only 6 of 149 calls were excluded due to unclear documentation. Therefore, 143 calls (96% of all applicable calls) were used to generate the findings of this quality improvement study. This corresponds to a study that provides a greater overview of the GIM-COVID-19 Long-Term Sequelae RACE line calls than one with a lower percentage of population representation. This is indicative of sampling validity, and it translates to a greater elimination of design or inclusion bias. Prejudice in this study was also greatly eliminated as tabulation was performed by an outside party with no preexisting relationships with the subjects of the studied RACE line calls.

This study had some limitations. First, the preexisting medical conditions of PCPs' patients were not taken into consideration when tabulating reported postinfection symptoms. Preexisting conditions may have exacerbated the prevalence of some symptoms in patients with long COVID-19. Second, there is no clear differentiation in the severity of symptoms described by PCPs regarding their patients and no standardized approach to measure these symptoms on RACE line calls. It is also not known if other specialist RACE lines were called for individuals with more specific organ system symptoms (eg, respirology, cardiology, psychiatry). Therefore, this study's data may reflect a smaller number of true RACE calls for long COVID. Third, the retrospective nature of the study limits our ability to verify or clarify PCP interpretations of patient symptoms. Additionally, given the absence of severity scoring or follow-up outcomes, we cannot correlate RACE call content with long-term patient trajectories. Nonetheless, our use of deidentified service data and a near-complete inclusion rate minimizes bias and supports the robustness of thematic trends observed.

Conclusion

The GIM-COVID-19 Long-Term Sequelae RACE line was introduced to provide PCPs with prompt medical advice from GIM specialists regarding chronic COVID. Over the course of this pandemic, this resource has been used by health care professionals to improve timely access to care, and it has provided support for the appropriate delivery of high-quality care by PCPs. This study investigated GIM-COVID-19 Long-Term Sequelae RACE line calls between August 2020 and June 2021, revealing many trends in the data. These calls mainly involved consults regarding the long COVID-19 symptoms being experienced by PCPs' patients. Respiratory symptoms were the leading type of symptom reported, with shortness of breath, cough, fatigue, and fever being the most common, respectively.

Moving forward, RACE calls can be monitored in situations of emerging diseases to better inform and educate community physicians about common complaints that patients are presenting with. The infrastructure and success of the RACE line display the value of creating specialty-informed, swift, and scalable support services during potential future public health emergencies. Integrating rapid-access consultation lines within more broad interdisciplinary networks, such as the PC-ICCN, can enable the timely translation of knowledge and reduce primary care uncertainty for atypical or complex conditions. The burden of chronic and multisystemic conditions is increasing, even outside the pandemic, and future iterations of RACE models could be modified to support conditions like myalgic encephalomyelitis/chronic fatigue syndrome, postviral syndromes, or multimorbidity in aging populations. Policymakers and health system leaders should consider sustained funding and integration of virtual consult models as part of long-term primary care innovation. The experience in British Columbia reflects a pattern seen internationally. Health systems across the United Kingdom, the United States, Australia, and several European countries have developed comparable virtual or rapid-access pathways to manage post-COVID. Placing the RACE line within this wider global effort highlights its importance not only as a local quality improvement initiative but also as a practical example of how specialist consultation models can support the evolving response to complex long COVID care.

Acknowledgments

The authors of this quality improvement study thank all primary care providers who used the COVID-GIM-Post-Infection Care RACE line, thereby providing the data that were tabulated in this study.

Authors' Contributions

SK, PB, JG, AL, MM, ZS conceived and designed the study. SK, PB, JG, AL, MM, ZS were responsible for data collection and initial analysis. JB further contributed to additional data analysis and interpretation, conducted a comprehensive review of relevant literature, and provided critical revisions that improved the overall quality of the manuscript, along with the other authors. All authors contributed to manuscript drafting and revision, approved the final version, and agree to be accountable for all aspects of the work.

Conflicts of Interest

None declared.

References

1. Rapid access to consultative expertise: an innovative model of shared care. RACE. 2022 Jan 6. URL: http://www.raceconnect.ca/about-race/what-is-race/ [accessed 2025-07-24]
2. Naming the coronavirus disease (COVID-19) and the virus that causes it. World Health Organization. 2022 May 12. URL: https://www.who.int/emergencies/diseases/novel-coronavirus-2019/technical-guidance/naming-the-coronavirus-disease-(covid-2019)-and-the-virus-that-causes-it [accessed 2025-07-24]
3. A clinical case definition of post COVID-19 condition by a Delphi consensus, 6 October 2021. World Health Organization. 2021 Oct 6. URL: https://www.who.int/publications/i/item/WHO-2019-nCoV-Post_COVID-19_condition-Clinical_case_definition-2021.1 [accessed 2025-07-24]
4. Ritchie H, Mathieu E, Rodés-Guirao L, Appel C, Giattino C, Ortiz-Ospina E, et al. Coronavirus pandemic (COVID-19). Our World in Data.: Oxford: Global Change Data Lab; 2022 May 12. URL: https://ourworldindata.org/coronavirus [accessed 2025-07-24]
5. BC COVID-19 data. BC Centre for Disease Control. 2022 May 12. URL: http://www.bccdc.ca/health-info/diseases-conditions/covid-19/data [accessed 2025-07-24]
6. Proal AD, VanElzakker MB. Long COVID or post-acute sequelae of COVID-19 (PASC): an overview of biological factors that may contribute to persistent symptoms. Front Microbiol 2021;12:698169. [doi: 10.3389/fmicb.2021.698169] [Medline: 34248921]
7. Global prevalence of post-acute sequelae of COVID-19 (PASC) or long COVID: a meta-analysis and systematic review. BMJ Open 2023 Feb;13(2):e065284. [doi: 10.1136/bmjopen-2022-065284]
8. Resendez S, Brown SH, Ruiz Ayala HS, et al. Defining the subtypes of long COVID and risk factors for prolonged disease: population-based case-crossover study. JMIR Public Health Surveill 2024 Apr 30;10:e49841. [doi: 10.2196/49841] [Medline: 38687984]
9. Long COVID signs and symptoms. Centers for Disease Control and Prevention. URL: https://www.cdc.gov/covid/long-term-effects/long-covid-signs-symptoms.html [accessed 2025-07-24]
10. Meza-Torres B, Delanerolle G, Okusi C, et al. Differences in clinical presentation with long COVID after community and hospital infection and associations with all-cause mortality: English Sentinel Network Database Study. JMIR Public Health Surveill 2022 Aug 16;8(8):e37668. [doi: 10.2196/37668] [Medline: 35605170]
11. Castanares-Zapatero D, Chalon P, Kohn L, et al. Pathophysiology and mechanism of long COVID: a comprehensive review. Ann Med 2022 Dec;54(1):1473-1487. [doi: 10.1080/07853890.2022.2076901] [Medline: 35594336]
12. NIH launches long COVID clinical trials through RECOVER initiative, opening enrollment. National Institutes of Health. 2023 Jul 31. URL: https://www.nih.gov/news-events/news-releases/nih-launches-long-covid-clinical-trials-through-recover-initiative-opening-enrollment [accessed 2025-07-24]
13. Reviewing RECOVER's impact in 2024. RECOVER COVID initiative. URL: https://recovercovid.org/news/reviewing-recovers-impact-2024 [accessed 2025-07-24]
14. Farrell A, O'Flynn J, Jennings A. An investigation into general practitioners' experience with long COVID. Ir J Med Sci 2024 Dec;193(6):2869-2873. [doi: 10.1007/s11845-024-03782-7] [Medline: 39162988]
15. About the PC-ICCN. Provincial Health Services Authority. 2025. URL: http://www.phsa.ca/our-services/programs-services/post-covid-19-care-network/about [accessed 2025-07-24]
16. About RACE. South Island Division of Family Practice. URL: https://divisionsbc.ca/south-island/race/about-race [accessed 2025-07-24]
17. Singh J, Quon M, Goulet D, Keely E, Liddy C. The utilization of electronic consultations (eConsults) to address emerging questions related to long COVID-19 in Ontario, Canada: mixed methods analysis. JMIR Hum Factors 2025 Feb 28;12:e58582. [doi: 10.2196/58582] [Medline: 40019816]
18. The NHS plan for improving long COVID services. NHS England. 2023 Jul. URL: https://www.england.nhs.uk/publication/the-nhs-plan-for-improving-long-covid-services/ [accessed 2025-07-24]
19. Allard N, Miller A, Morgan M, Chakraborty S. Post-COVID-19 syndrome/condition or long COVID: persistent illness after acute SARS CoV-2 infection. Aust J Gen Pract 2022 Dec;51(12):952-957. [doi: 10.31128/AJGP-05-22-6429] [Medline: 36451331]
20. Parkin A, Davison J, Tarrant R, et al. A multidisciplinary NHS COVID-19 service to manage post-COVID-19 syndrome in the community. J Prim Care Community Health 2021;12:21501327211010994. [doi: 10.1177/21501327211010994] [Medline: 33880955]
21. Greenhalgh T, Darbyshire JL, Lee C, Ladds E, Ceolta-Smith J. What is quality in long covid care? Lessons from a national quality improvement collaborative and multi-site ethnography. BMC Med 2024 Apr 15;22(1):159. [doi: 10.1186/s12916-024-03371-6] [Medline: 38616276]

22. Brigham E, O'Toole J, Kim SY, et al. The Johns Hopkins Post-Acute COVID-19 Team (PACT): a multidisciplinary, collaborative, ambulatory framework supporting COVID-19 survivors. Am J Med 2021 Apr;134(4):462-467. [doi: 10.1016/j.amjmed.2020.12.009] [Medline: 33444589]
23. Whyler N, Atkins L, Hogg P, et al. Harnessing the benefits of telehealth in long COVID service provision. Public Health Rev 2024;45:1606948. [doi: 10.3389/phrs.2024.1606948] [Medline: 38881555]
24. Floridia M, Grassi T, Giuliano M, et al. Characteristics of long-COVID care centers in Italy. A national survey of 124 clinical sites. Front Public Health 2022;10:975527. [doi: 10.3389/fpubh.2022.975527] [Medline: 36062113]

Abbreviations

GIM: general internal medicine
GP: general practitioner
PASC: postacute sequelae of SARS-CoV-2
PC-ICCN: Post-COVID-19 Interdisciplinary Clinical Care Network
PCP: primary care practitioner
RACE: Rapid Access to Consultative Expertise

Edited by A Schwartz; submitted 02.Feb.2024; peer-reviewed by SO Olalere, Anonymous; revised version received 24.Oct.2025; accepted 29.Oct.2025; published 10.Feb.2026.

Please cite as:
Kaushal S, Bhandal J, Birks P, Greiner J, Levin A, Malbeuf M, Schwartz Z
Use of a Specialist Telephone Consultation Line for Long COVID in Primary Care in British Columbia: Retrospective Descriptive Quality Improvement Study
JMIRx Med 2026;7:e57021
URL: https://xmed.jmir.org/2026/1/e57021
doi:10.2196/57021

The Performance of DeepSeek R1 and Gemini 3 in Complex Medical Scenarios: Comparative Study

Maria Bajwa[1*], PhD; Robert Hoyt[2*], MD; Dacre Knight[3*], MD; Maruf Haider[4], MD

[1]MGH Institute of Health Professions, Boston, MA, United States

[2]Internal Medicine Department, Virginia Commonwealth University, 57 North 11th Street, Richmond, VA, United States

[3]Internal Medicine Department, University of Virginia, Charlottesville, VA, United States

[4]Internal Medicine Department, Carilion Roanoke Memorial Hospital, Roanoke, VA, United States

[*]these authors contributed equally

Corresponding Author:
Robert Hoyt, MD
Internal Medicine Department, Virginia Commonwealth University, 57 North 11th Street, Richmond, VA, United States

Related Articles:

Companion article: https://www.medrxiv.org/content/10.1101/2025.04.29.25326666v1

Companion article: https://med.jmirx.org/2026/1/e96223

Companion article: https://med.jmirx.org/2026/1/e96225

Companion article: https://med.jmirx.org/2026/1/e96227

Companion article: https://med.jmirx.org/2026/1/e96220

Abstract

Background: Generative artificial intelligence models, especially reasoning large language models (LLMs), are gaining adoption in health care for diagnostic decision support and medical education. DeepSeek R1 is a reasoning LLM that generates extended chain-of-thought explanations to make its decision-making process more explicit. Traditional medical benchmarks often lack complexity and authenticity, motivating the adoption of scenario-rich datasets, such as the Massive Multitask Language Understanding Pro (MMLU-Pro) professional medicine subset, which provides multispecialty clinical vignettes for reasoning-centric evaluation.

Objective: The objective of this study is to assess the diagnostic accuracy, reasoning quality, reasoning transparency, and practical usability of DeepSeek R1 and Gemini 3 Pro across closed- and open-ended clinical scenarios, with the intention of guiding their prospective application in practical clinical education and training. This evaluation was conducted by analyzing 162 diverse medical scenarios (both closed- and open-ended) from the MMLU-Pro health subset.

Methods: In a 2-phase, dual-model evaluation, DeepSeek R1 and Gemini 3 Pro were applied to 162 matched clinical vignettes from the MMLU-Pro professional medicine subset spanning 21 specialties. Closed-ended, multiple-choice, and open-ended prompts were constructed for the same scenarios, and model outputs were coded for accuracy, reasoning steps, and citation behavior; descriptive statistics and the McNemar test were used to compare performance across formats.

Results: DeepSeek R1 achieved an accuracy of 86.4% (140/162 scenarios) on closed-ended tasks and 80.9% (131/162) on open-ended questions across 162 clinical scenarios, indicating modest attenuation of performance when answer cues were removed. Gemini 3 Pro demonstrated 90.7% (147/162) closed-ended and 88.9% (144/162) open-ended accuracy on the same scenarios, showing a similar pattern of decreased performance without answer options. Error analysis indicated that incorrect answers typically involved longer reasoning chains, suggesting overthinking. In a structured review of open-ended responses, DeepSeek R1 produced an average of 18.7 (range 0 - 52) references per case, with 5.2 unrelated references and 13.1 (range 3 - 67) reasoning steps, whereas Gemini 3 Pro averaged 22.5 (range 12 - 50) references, 1.9 (range 0 - 8) unrelated references, and 4.4 (range 1 - 10) reasoning steps per case.

Conclusions: DeepSeek R1 demonstrated moderate-to-excellent accuracy and reasoning in evaluating both closed- and open-ended medical scenarios. In parallel, Gemini 3 Pro showed broadly comparable but distinct performance and reasoning patterns. While

the closed-ended format may inflate accuracy due to cueing, the open-ended evaluation yielded richer insights into the fidelity of reasoning. Side-by-side evaluation of two large reasoning models highlights the importance of format, specialty, and citation behavior when considering clinical and educational use. Continued validation across a wider range of specialties and real-world contexts will enhance the model's trustworthiness for diagnostic and teaching applications.

(JMIRx Med 2026;7:e76822) doi:10.2196/76822

KEYWORDS

large reasoning model; LRM; large language model; LLM; accuracy; medical scenario; DeepSeek R1; Gemini 3

Introduction

Generative artificial intelligence (AI) models, particularly large language models (LLMs), have demonstrated substantial advancements across various health care domains, including diagnostics, patient management, clinical documentation, and medical education [1,2]. With the emergence of reasoning-focused architectures, such as DeepSeek R1, the paradigm has shifted from text prediction to structured inference, characterized by chain-of-thought reasoning, mixture of experts, reinforcement learning, and more transparent decision paths [3-5]. Recent work has highlighted DeepSeek R1 as an open-source reasoning LLM with visualized decision pathways and low-cost deployment, and it has garnered growing interest in clinical decision support, patient engagement, and medical education; nevertheless, researchers have emphasized ongoing challenges related to hallucinations, modality limitations, and ethical integration into health care systems [6]. The features of reasoning LLMs are activated iteratively to answer a user's zero-shot prompt, enabling clinicians and educators to use reasoning models in simple, instruction-based interactions [4,5,7]. In parallel, Gemini 3 Pro (Google Inc) has emerged as a state-of-the-art multimodal LLM that integrates strong language reasoning with image and structured-data understanding and has demonstrated high performance on broad academic benchmarks, including Massive Multitask Language Understanding Pro (MMLU-Pro), as well as medical examination–style question sets such as Medical Question Answering (MedQA), based on the United States Medical Licensing Examination (USMLE) [8]. Public benchmarks and technical reports describe Gemini 3 Pro (and related Med-Gemini variants) achieving competitive or superior accuracy to prior models across general knowledge and health care–oriented tasks, while also highlighting ongoing concerns about hallucinations, transparency, and responsible clinical deployment, similar to other frontier systems [8]. To our knowledge, no prior work has reported open-ended diagnostic performance and reasoning metrics for Gemini 3 Pro on the MMLU-Pro professional medicine benchmark subset.

Even with these changes, the use of LLMs in the real world still raises concerns about reliability, bias, replicability, and generalizability. Traditional evaluation benchmarks, most notably MedQA/USMLE multiple-choice questions (MCQs), have facilitated initial assessments of model performance, yet they are increasingly criticized for plateauing scores, susceptibility to cueing, testwiseness effects, and limited specialty representation. In these examination-style settings, models can sometimes infer correct answers from partial cues or test-taking strategies rather than demonstrating robust clinical reasoning, potentially overestimating their readiness for real-world use [9-12]. To address these limitations, MMLU-Pro, a modification of the original MMLU benchmark, was developed as a more robust and challenging multitask language dataset [13,14]. The MMLU-Pro professional medicine (health) subset provides scenario-based clinical vignettes across multiple specialties and is designed to increase diagnostic reasoning complexity, reduce test-taking artifacts, and broaden domain coverage compared with earlier examination-style benchmarks. A recent systematic review of 39 medical LLM benchmarks further quantified that examination-style "knowledge-based" benchmarks often report high accuracies (approximately 84%-90%), whereas more practice-based, clinically oriented benchmarks show substantially lower performance (approximately 45%-69%), particularly for clinical reasoning and safety, underscoring a persistent knowledge-and-practice performance gap and the need for richer, scenario-focused evaluation frameworks [15].

It is believed that the MMLU-Pro health subset increases the complexity of diagnostic reasoning, reduces test-taking artifacts, and provides a broader domain representation [13]. Previous research has evaluated the questioning capabilities of LLMs [16]. Much of this work has relied on single-format multiple-choice or short-answer medical question-answering benchmarks, emphasizing aggregate accuracy on examination-style items rather than detailed analyses of clinical reasoning processes [16-18]. Nonetheless, limited research has used MMLU-Pro in a scenario-rich, dual-format assessment that encompasses both constrained (closed-ended) and expressive (open-ended) reasoning tasks. In this context, closed-ended tasks require the model to select an answer from predefined options (for example, choosing a single best diagnosis from 4 MCQ choices), whereas open-ended tasks require the model to generate a free-text diagnosis and supporting reasoning without explicit answer cues, more closely approximating how clinicians articulate and justify diagnostic judgments in practice. [15-18,] The MMLU-Pro dataset is unique in that it offers 10 potential answers for a model to choose from. Based on prior work, we hypothesized that the question format of open- and closed-ended questions would have a meaningful impact on model performance in complex medical scenarios and clinical reasoning [17,18].

We align our study with the Transparent Reporting of a Multivariable Model for Individual Prognosis or Diagnosis—Large Language Model (TRIPOD-LLM) guidelines to offer new insights into model behavior across diverse clinical tasks [19]. TRIPOD-LLM guidance ensured a structured

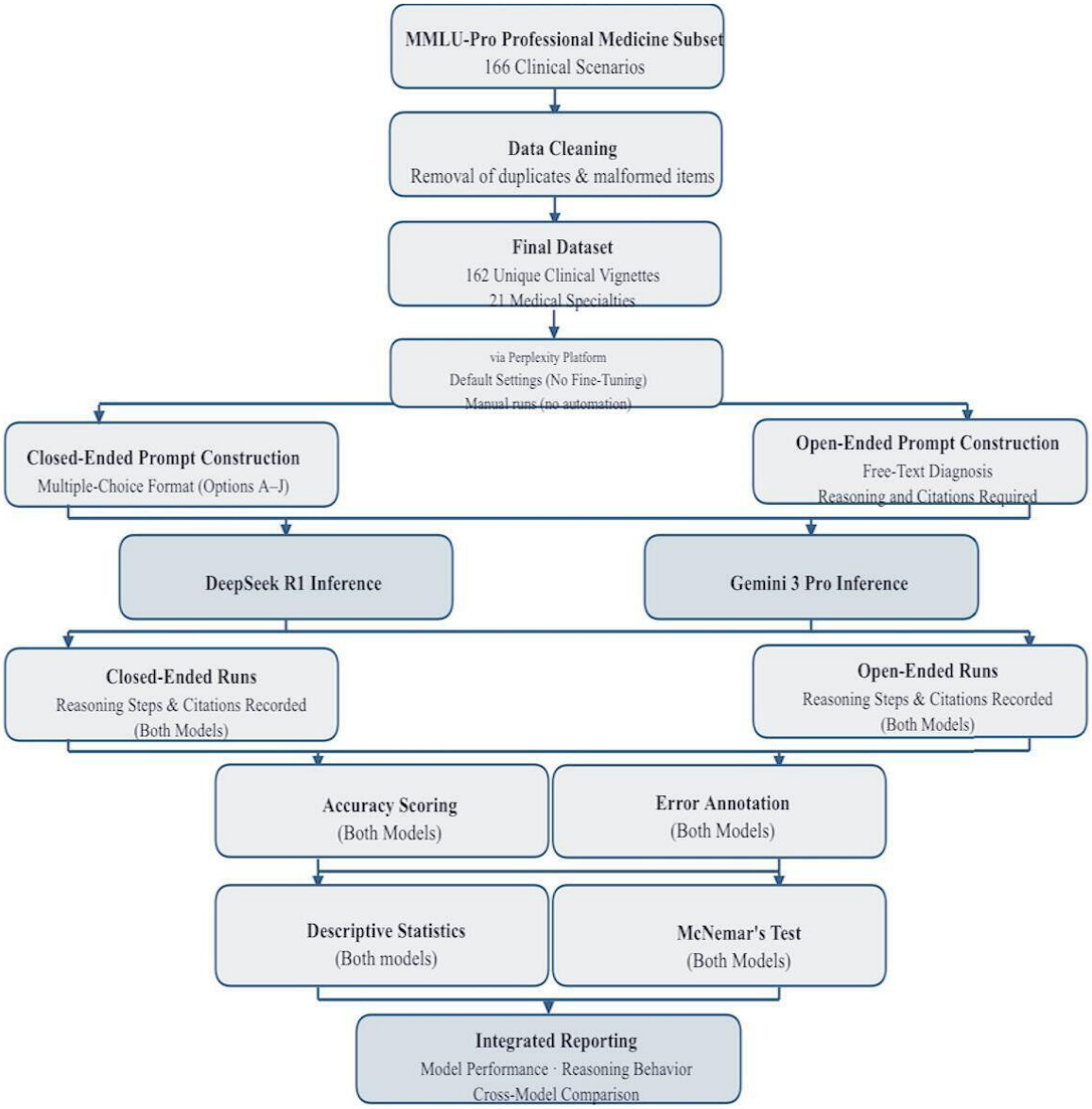
protocol to highlight transparency, reproducibility, and standardized reporting across all evaluation stages [19]. The intended audience for this work includes clinicians, educators, and AI researchers interested in using or evaluating LLMs or large reasoning models (LRMs) for diagnostic decision support, curriculum design, and medical reasoning education. A completed TRIPOD-LLM compliance checklist, mapping each reporting item to corresponding manuscript sections, is provided in Multimedia Appendix 1.

The objective of this study is to assess the diagnostic accuracy and reasoning quality of DeepSeek R1 and Gemini 3 Pro with the intention of guiding their prospective application in practical clinical education and training. Gemini 3 was selected as a comparable LLM. More details on this model are provided in the Methods section. We aimed to (1) compare model accuracy and behavior across prompt formats using structured evaluation criteria, (2) assess variability in performance across clinical domains, and (3) characterize types of reasoning and interpretability errors to identify limitations in clinical reasoning and generalizability for both models.

Methods

We conducted a 2-phase, protocol-driven, 2-model exploratory evaluation of DeepSeek R1 and Gemini 3 Pro to test their performance and diagnostic behaviors across structured clinical vignettes [14]. (Figure 1)

Figure 1. Illustration of the study design, with 2 large language models.



Data Source and Preparation

The MMLU-Pro dataset contains over 12,000 questions in 14 categories [13]. We identified the subcategory of "professional medicine," comprising 166 complex medical scenarios categorized across 21 specialties and spanning a spectrum of real-world clinical contexts, from primary care to subspecialty practice. Duplicates and erroneously formatted cases were removed. Duplicates were defined as items with identical clinical stems and answer keys and were collapsed to a single representative scenario. Erroneously formatted cases were those with obvious structural problems (eg, missing answer options, incomplete vignette text, or clearly mismatched options in the MMLU-Pro files) and were excluded. A final set of 162

scenarios served as the basis for both study phases, each presented as a clinical vignette to assess diagnostic reasoning. We regarded the provided key to the answers as the definitive source of information. For closed-ended items, responses were scored as correct if the selected letter matched the keyed answer. For open-ended items, free-text diagnoses were considered correct if they exactly matched or were clear clinical synonyms of the keyed diagnosis (eg, “acute myocardial infarction” vs “myocardial infarction”), as determined by author review. Three practicing clinicians on the author team independently resolved any ambiguous cases by consensus discussion. No imputation was performed; error breakdowns are reported for all scenarios, and the dataset was last modified in June 2024. An example scenario is shown in Textbox 1.

Textbox 1. Example scenario (#6037).

A 47-year-old man is brought to the emergency department 2 hours after the sudden onset of shortness of breath, severe chest pain, and sweating. He has no history of similar symptoms. He has hypertension treated with hydrochlorothiazide. He has smoked one pack of cigarettes daily for 30 years. His pulse is 110/min, respirations are 24/min, and blood pressure is 110/50 mm Hg. A grade 3/6 diastolic blowing murmur is heard over the left sternal border and radiates to the right sternal border. Femoral pulses are decreased bilaterally. An ECG shows left ventricular hypertrophy. Which of the following is the most likely diagnosis? ['A. Acute myocardial infarction' B. 'Congestive heart failure' C. 'Angina pectoris' D. 'Aortic dissection' E. 'Mitral valve prolapse' F. 'Esophageal rupture' G. 'Hypertensive crisis' H. 'Thoracic aortic aneurysm' I. 'Pulmonary embolism' J. 'Aortic stenosis']

Models Evaluated

The evaluation was conducted on DeepSeek R1 (January 2025 release), a publicly available, open-source LRM [20]. We used an uncensored version of DeepSeek hosted by a well-known AI search engine platform, Perplexity AI (Perplexity AI, Inc) [21,22]. No model fine-tuning, postprocessing, or temperature modification was applied. No post hoc calibration, bias correction, or output pruning was performed.

We also evaluated Gemini 3 Pro, a proprietary large multimodal model with advanced reasoning capabilities (version 1, released November 18, 2025) [23]. The Gemini 3 series excels at complex tasks involving text, images, video, audio, and code. Technical features include a 1-million-token context window and native multimodal support. The models support up to 3000 images or 45 minutes of video in a single prompt. The knowledge cutoff date is January 2025 [23]. Gemini 3 Pro was accessed through the same conversational interface using identical prompts, scenarios, and default settings, with no fine-tuning, temperature modification, or post hoc calibration. Both models were run manually (no automation), one scenario per prompt, for closed- and open-ended questions.

Computational Resources

Both DeepSeek R1 and Gemini 3 Pro were accessed via Perplexity AI’s web chat platform (model version not explicitly labeled; presumed latest release at the time of evaluation). All runs used the same conversational interface, prompts, and default model settings, with no fine-tuning, temperature, or max-token modification, post hoc calibration, bias correction, or output pruning. Inference was executed on Perplexity’s standard cloud infrastructure, with typical per-query latency of approximately 4 to 20 seconds; detailed hardware configuration, server location, and floating-point throughput are unavailable. DeepSeek R1 was evaluated from March 6 to 10, 2025, for closed-ended questions, and from March 12 to 15, 2025, for open-ended questions, using default settings (temperature 1.0, max tokens 2048) and one scenario per prompt, as batch execution was not available. Gemini 3 Pro was evaluated on the same 162 scenarios under identical manual execution procedures, from January 17 to 22, 2026, for closed-ended questions and February 12 to 25, 2026, for open-ended questions, again with one scenario per prompt and no automation agent.

Evaluation Protocol

The evaluation protocol closely follows TRIPOD-LLM’s recommendations for rigorous, open LLM evaluation in health contexts and is reported accordingly [19]. The study evaluated the performance of DeepSeek R1 and Gemini 3 on the MMLU-Pro professional medicine subset across 2 phases, using both closed-ended MCQs and open-ended questions [20-23]. To interact with the LRM, prompts were designed with clear, reproducible instructions; in the open-ended format, the only modification was removing answer choices and phrasing cues. The full 2-phase protocol was then repeated with Gemini 3 Pro using the same prompts, scenarios, and execution procedures.

In both phases and both LRMs, disagreements with the MMLU-Pro key were recorded and reported without further resolution for future review. All queries were run manually without the aid of an automation agent. A structured error-annotation system, described in the Qualitative Analysis section, was applied to all incorrect answers to characterize model behavior. Relatedness, nonalignment to the answer keys, or hallucination of citations were coded by the principal investigator and reviewed for consensus among the author team.

Closed-Ended Phase

The model’s chosen answer (A-J) and generated rationale and literature citations were recorded for each scenario. We measured the actual reasoning steps for closed- and open-ended questions. Model answers were compared to the official MMLU-Pro answer key. For each incorrect response, an error category was then assigned using the structured taxonomy described below, and the associated citation count and reasoning-step count were recorded for subsequent descriptive analysis.

The closed-ended prompt was as follows: “As a medical consultant, you will respond to questions about various medical scenarios. This is a multiple-choice question with up to 10 options. Select the answer that is most likely. Report the correct answer A-J. Then, provide your reasoning steps and cite relevant literature.” The input structure had (1) the prompt text, (2) a full scenario vignette, and (3) a list of answer choices (A, B,… J).

Open-Ended Phase

In the second phase, scenarios were reformatted as open-ended prompts that required diagnosis and differential diagnosis without provided options. For each item, the original MMLU-Pro vignette stem was preserved verbatim, while the last clause, "Which of the following is the most likely diagnosis?" (or equivalent) and the accompanying answer list were removed so that the model received the same clinical information in both phases but without explicit options in the open-ended condition. Both DeepSeek R1 and Gemini 3 Pro were prompted with the open-ended version using the following prompt: "As a senior clinician, you have received requests for consultation on various medical scenarios. Provide a diagnosis and differential diagnosis for each case. Then, explain your reasoning and cite relevant literature." The input structure included (1) the above prompt text, (2) a full scenario vignette, and (3) no answer options.

The output included diagnosis, differentials, rationale, citations, and possible hallucinations and was recorded. Outputs were assessed by 2 board-certified clinician authors (MH and DK), each reviewing half the cases for accuracy and alignment with the ground truth, and discrepancies were resolved by a third clinician author (RH). Afterwards, the same structured error-annotation framework used in the closed-ended phase was used to assign the errors in this phase. Each incorrect open-ended response was assigned to a single error category, and its citation count and reasoning-step count were extracted using the same procedures as for closed-ended items.

Statistical Analysis

Descriptive statistics, including frequency and proportion, were calculated for model accuracy, citation count, number of reasoning steps, error frequencies, and specialty-specific performance. For both formats, each scenario was scored as correct (1) or incorrect (0) based on a comparison with the MMLU-Pro answer key, as described above, and overall accuracy was calculated as the proportion of correctly answered items out of 162. Citation counts were obtained by manually counting the distinct reference links or source objects produced by the models in each response. The number of reasoning steps was taken from the model's own structured reasoning output (eg, "Step 1… Step 2…"), using the step count reported in the response when available and, when absent, by counting discrete reasoning statements in the model's explanation. Each vignette was mapped to a single specialty using the professional medicine labels in MMLU-Pro, supplemented by manual assignments when necessary, to derive specialty-specific accuracy. Each scenario was scored as correct or incorrect, and model accuracy was compared between closed- and open-ended formats. The McNemar test was then applied to the paired binary data of closed- vs open-ended questions for each model to determine whether there was a statistically significant difference in results between the 2 related groups, as recommended in the statistical literature for within-subject categorical comparisons [24,25]. No formal statistical comparisons were conducted between formats for citation or reasoning metrics due to differences in sampling completeness and the limited number of questions. All statistical tests were 2-tailed, and results were considered significant at $P<.05$. No power calculation was performed due to the fixed number of scenarios (n=162).

Qualitative Analysis

To categorize the errors, we developed a structured taxonomy for the error annotation system by analyzing incorrect outputs thematically. All queries were executed manually without the use of automated agents. The error annotation framework was informed by prior literature [26] and refined through consensus among the research team. Initial error categories were drafted based on observed response patterns and iteratively revised to ensure consistent interpretation. All incorrect outputs were reviewed using the original dataset, including reasoning steps, citation counts, and specialty labels, and disagreements were resolved through team discussion.

Ethical Considerations

No patient or human subject data were used. Therefore, approval from the ethical review board was not required. Model, prompt, and grading protocols are available upon request.

Results

Model Performance

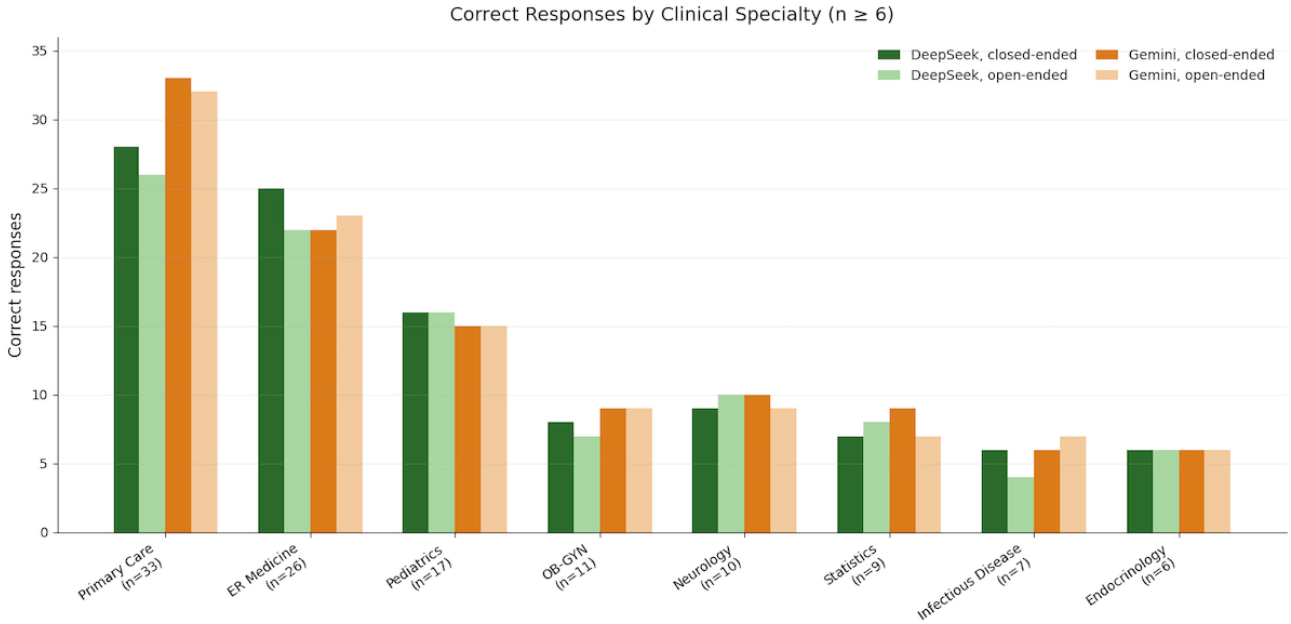
DeepSeek R1 and Gemini 3 Pro were each evaluated on 162 clinical scenarios from the MMLU-Pro professional medicine subset spanning multiple medical specialties. Both models were assessed on the same cases using closed-ended (multiple-choice) and open-ended (free-response) formats. The largest specialty groups included primary care, emergency medicine, pediatrics, obstetrics and gynecology, and neurology, each contributing 10 or more cases. Model responses were classified as correct or incorrect to assess performance across question formats and clinical domains. Specialty-level distribution and diagnostic accuracy for specialties with 6 or more scenarios are summarized in Table 1; the full set of specialties is provided in Multimedia Appendix 2.

Table . Diagnostic accuracy by clinical specialty for DeepSeek R1 and Gemini 3 Pro across closed- and open-ended formats for specialties with 6 cases or more in the question set. The last row provides an average per column for the 8 specialties.

Clinical specialty	DeepSeek R1 (n=162)		Gemini 3 Pro (n=162)	
	Closed-ended correct, n (%)	Open-ended correct, n (%)	Closed-ended correct, n (%)	Open-ended correct, n (%)
Primary care (n=33)	28 (84.8)	26 (78.8)	33 (100.0)	32 (97.0)
Emergency medicine (n=26)	25 (96.2)	22 (84.6)	22 (84.6)	23 (88.5)
Pediatrics (n=17)	16 (94.1)	16 (94.1)	15 (88.2)	15 (88.2)
Obstetrics and gynecology (n=11)	8 (72.7)	7 (63.6)	9 (81.8)	9 (81.8)
Neurology (n=10)	9 (90.0)	10 (100.0)	10 (100.0)	9 (90.0)
Statistics (n=9)	7 (77.8)	8 (88.9)	9 (100.0)	7 (77.8)
Infectious disease (n=7)	6 (85.7)	4 (57.1)	6 (85.7)	7 (100.0)
Endocrinology (n=6)	6 (100.0)	6 (100.0)	6 (100.0)	6 (100.0)
Average score for 8 specialties, %	87.8	83.5	92.6	90.5

DeepSeek R1 answered 140 of 162 scenarios correctly in the closed-ended format (86.4%), whereas Gemini 3 Pro answered 147 (90.7%). Accuracy was generally high across clinical domains, with moderate variation by specialty (Table 1). When answer options were removed, performance declined for both models: DeepSeek R1 answered 131 scenarios correctly (80.9%), and Gemini 3 Pro answered 144 (88.9%). The reduction in accuracy associated with removing answer options was greater for DeepSeek R1 than for Gemini 3 Pro (5.5 vs 1.8 percentage points). Accuracy varied across specialties, with obstetrics and gynecology demonstrating the lowest average accuracy (75.2%) (Table 1; Figure 2).

Figure 2. Diagnostic accuracy of DeepSeek R1 and Gemini 3 Pro across clinical specialties with ≥6 evaluated scenarios.



Paired Accuracy Comparison

To assess whether diagnostic accuracy differed between question formats, the McNemar test was applied to paired responses for each model across the 162 scenarios. For DeepSeek R1, the difference between closed-ended and open-ended accuracy was not statistically significant (χ^2_1=3.37; *P*=.07). Similarly, Gemini 3 Pro showed no significant difference between formats (χ^2_1=0.21; *P*=.65). These findings indicate that removing answer options did not produce a statistically significant change in diagnostic accuracy for either model.

References and Reasoning Steps

Table 2 demonstrates that, in the open-ended condition, DeepSeek R1 included more references and reasoning steps than Gemini 3 Pro but had lower accuracy. One response (Q#6111) represented an extreme outlier, containing 70 unrelated references, largely focused on HIV literature rather than the statistical reasoning required by the scenario. Citation and reasoning metrics are reported for open-ended responses only (n=162 per model); closed-ended citation and reasoning data were not collected comprehensively for both models. Because our primary interest was in understanding diagnostic

reasoning behavior in unconstrained, real-world–like settings, we prioritized coding of citations and reasoning steps for open-ended responses, whereas closed-ended outputs were evaluated primarily for accuracy and error patterns rather than detailed citation analysis.

Table . References, unrelated references, and reasoning steps per model on open-ended questions.

	References		Unrelated references	Steps	
Model	Mean (range)	Median	Unrelated references, mean (range)	Mean (range)	Median
DeepSeek	33.1 (0 - 84)	30	3.1 (0 - 70)	10.7 (3 - 34)	10
Gemini	22.5 (12 - 50)	15	1.9 (0 - 8)	4.4 (1 - 10)	4

Error Analysis

Errors were unevenly distributed across specialties and question formats. For DeepSeek R1, closed-ended errors were concentrated in primary care (n=5; questions #6249, #6432, #6493, #6501, #6507), obstetrics and gynecology (n=3; questions #6183, #6307, #6664), and general surgery (n=3; questions #6491, #6503, #6611). In the open-ended condition, errors increased in several of these specialties and also appeared in emergency medicine (n=4; questions #6310, #6544, #6748, #6793), infectious disease (n=3; questions #6236, #6242, #6682), and psychiatry (n=2; questions #6497, #6674), specialties that had few or no errors in the closed-ended format. Overall, DeepSeek R1 produced more errors in the open-ended format (n=31) than in the closed-ended format (n=22), with 8 specialties showing reduced accuracy when answer options were removed.

For Gemini 3 Pro, the distribution differed. Closed-ended errors occurred in emergency medicine (n=3; questions #6310, #6557, #6796), pediatrics (n=2; questions #6252, #6430), obstetrics and gynecology (n=2; questions #6183, #6307), general surgery (n=2; questions #6503, #6608), and psychiatry (n=2; questions #6112, #6674). In contrast to DeepSeek R1, the open-ended format improved performance in several of these areas: 8 questions answered incorrectly in the closed-ended condition were answered correctly in the open-ended condition. New open-ended errors appeared primarily in specialties that had no closed-ended failures, including oncology (n=2; questions #6187, #6618), statistics (n=2; questions #6111, #6664), and trauma surgery (n=1; question #6043). Three specialties—endocrinology, nephrology, and immunology—showed no errors across either model or format.

Persistent errors further clarified differences between models. DeepSeek R1 answered 17 items incorrectly in both formats (questions #6045, #6183, #6184, #6249, #6252, #6307, #6310, #6432, #6491, #6493, #6501, #6503, #6550, #6611, #6664, #6672, #6682), spanning primary care, obstetrics and gynecology, general surgery, and infectious disease. These persistent errors represented 77% (17/22) of DeepSeek R1's closed-ended failures and 55% (17/31) of its open-ended failures, suggesting that many of its mistakes reflected format-independent uncertainty rather than sensitivity to question type.

Gemini 3 Pro demonstrated fewer persistent errors (n=7; questions #6252, #6307, #6503, #6550, #6557, #6674, #6796), concentrated mainly in emergency medicine, obstetrics and gynecology, and psychiatry. These accounted for 47% of its closed-ended failures, indicating that Gemini's performance was more responsive to question type. Four questions (#6252 [pediatrics], #6307 [obstetrics and gynecology], #6503 [general surgery], and #6550 [vascular surgery]) were answered incorrectly by both models across both formats, suggesting question-level complexity or ambiguity rather than model-specific limitations (Table 3).

Table . Error taxonomy with representative examples for DeepSeek R1 and Gemini 3 Pro. Error taxonomy was applied to all incorrect open-ended responses for both models and to incorrect closed-ended responses for DeepSeek R1. "Not observed" indicates that the error type was not identified in any response for that model.

Error type	Description	DeepSeek R1 example	Gemini 3 Pro example
E1: Reasonable but nonoption answer	Clinically valid response that diverges from the benchmark key	Question #6118 (primary care): model provided a plausible but unlisted management response in the open-ended format; correct in the closed-ended condition (both closed and open formats).	Question #6043 (trauma surgery): model selected a clinically defensible approach not aligned with the MMLU-Pro answer key (both closed and open formats).
E2: Overthought/ contradictory reasoning	Excessively long or internally inconsistent reasoning chain	Question #6491 (general surgery): closed-ended response generated 56 reasoning steps before reaching an incorrect conclusion, with internal contradictions across steps (both closed and open formats).	Not observed.
E3: Citation hallucination	Fabricated or unverifiable citations presented as real references	Not observed.	Not observed.
E4: Missing differentials	Failure to address relevant diagnostic alternatives	Question #6672 (allergy immunology): open-ended response did not consider AM cortisol measurement as a diagnostic step, which was central to the correct answer (both closed and open formats).	Question #6104 (neurology): open-ended response failed to incorporate key neurological differentials, leading to an incomplete diagnostic approach (both closed and open formats).
E6: Circular logic	Repetition of reasoning without progress toward diagnosis	Question #6310 (emergency medicine): closed-ended response generated 31 reasoning steps that restated the clinical scenario repeatedly without narrowing the differential or reaching a defensible conclusion (both closed and open formats).	Not observed.
E7: Ambiguous/overloaded output	Diffuse reasoning with excessive or tangential information	Question #6254 (genetics): closed-ended response produced 46 citations for a genetics question, with tangential references diluting the clinical focus of the answer (both closed and open formats).	Not observed.

In several cases across both models, responses classified as incorrect were judged by the authors to represent clinically reasonable alternatives, reflecting updated guidelines or equivalent management strategies not captured by the MMLU-Pro answer key. This pattern was particularly prominent for Gemini 3 Pro, where 17 of 18 open-ended errors (94%) were classified as E1 (reasonable but nonoption answer). These findings suggest that benchmark accuracy alone may not fully capture clinically valid reasoning in LLMs.

Qualitative Analysis of the Errors

To understand the mechanisms underlying incorrect responses, we applied the structured error taxonomy described in the Methods section to all incorrect open-ended responses from both models and to closed-ended incorrect responses from DeepSeek R1. Each error was categorized into one of six predefined types: reasonable but nonoption answer (E1), overthought or contradictory reasoning (E2), citation hallucination (E3), missing differentials (E4), circular logic (E6), and ambiguous or overloaded output (E7) (Table 2).

Across the dataset, the most common error type was E1 (reasonable but nonoption answer), followed by E2 (overthought or contradictory reasoning) and E4 (missing differentials). Citation hallucination (E3) was not observed in either model, indicating that the models' references were generally grounded in real sources. Overthought reasoning and ambiguous outputs (E2 and E7) were characterized by unusually long or diffuse reasoning chains. In contrast, E4 (missing differentials) occurred exclusively in open-ended responses for both models, consistent with the unconstrained prompt structure, allowing broader diagnostic elaboration without predefined answer options.

The two models exhibited distinct error profiles. DeepSeek R1 demonstrated a broader range of error types, including E1, E2, E4, E6, and E7. Several of these errors reflected instability in reasoning, such as excessively long chains of reasoning (E2) or circular diagnostic logic (E6). In contrast, Gemini 3 Pro's errors were overwhelmingly classified as E1, with 17 of 18 open-ended errors representing clinically plausible answers that diverged from the benchmark key. Only a single Gemini response showed a missing-differential error (E4). These contrasting patterns suggest that the models differed not only in overall accuracy but also in their failure modes: DeepSeek R1's errors more often reflected reasoning instability under uncertainty, whereas Gemini 3 Pro more frequently produced clinically plausible

answers that diverged from the benchmark answer key (Table 3).

Discussion

Principal Findings

Our evaluation demonstrated that Gemini 3 Pro achieved higher overall performance (90.7% closed-ended; 88.9% open-ended) than DeepSeek R1 (86.4% closed-ended; 80.9% open-ended). Accuracy declined modestly when answer options were removed for both models; however, paired analysis showed that this difference was not statistically significant.

DeepSeek R1 produced longer reasoning chains, but these did not improve accuracy. Its errors were distributed across multiple reasoning categories, including overthought reasoning, circular logic, and missing differentials. In contrast, most Gemini 3 Pro errors, particularly in open-ended responses, were classified as clinically reasonable answers. Consistent with prior evaluations of medical language models, a substantial proportion of incorrect responses in this study represented clinically plausible alternatives rather than clearly incorrect reasoning [9,12].

Specialty-level analysis demonstrated moderate variability in model performance across domains. Both models performed consistently well in pediatrics, neurology, and endocrinology, whereas lower accuracy was observed in specialties such as obstetrics and gynecology and primary care. Persistent errors across both question formats were relatively limited but suggest that certain scenarios posed intrinsic reasoning challenges independent of question type. The number of questions per specialty was small, limiting interpretation and generalizability. Accuracy reported here was consistent with, or exceeded, previous reports on prominent LLMs on medical benchmarks, including DeepSeek R1 and Gemini, as well as GPT-4, MedPaLM, and Claude [26-29].

Our results also contribute to a growing body of research examining how chain-of-thought reasoning affects model performance. While reasoning transparency is often considered a strength of LRMs, recent studies suggest that longer reasoning chains do not necessarily improve accuracy [30,31]. The reasoning patterns observed in DeepSeek R1 responses support this observation.

These results have implications for educational integration, a prominent opportunity for LRMs. When used with instructor oversight, reasoning models may support case-based learning, differential diagnosis exercises, and discussions of diagnostic reasoning processes. This transparency enables clinicians and educators to review diagnostic logic, reasoning chains, differentials, and literature-supported explanations as practical case-based learning resources [32,33].

Benchmark accuracy alone provides an incomplete assessment of a model's capabilities. For example, Gemini 3 Pro achieved higher accuracy with shorter reasoning outputs, suggesting that concise reasoning structures may sometimes reflect more stable inference processes. From a clinical safety perspective, the results reinforce current guidance that LLMs should not be used as independent diagnostic decision systems. Even highly accurate models occasionally produce reasoning errors or plausible but incorrect conclusions. Current expert recommendations emphasize that AI-generated clinical reasoning should be interpreted as decision-support information rather than authoritative clinical guidance and should always be reviewed by qualified clinicians.

Strengths and Limitations

This study has several strengths. First, it evaluated 2 LRMs using the same dataset, prompts, and evaluation protocol, allowing direct comparison of model behavior across identical clinical scenarios. Second, the dual-format design, incorporating both closed-ended and open-ended prompts, enabled assessment of diagnostic reasoning with and without answer cueing. This technique provides a more realistic evaluation of free-text clinical reasoning. The analysis integrated quantitative performance metrics with a structured qualitative error taxonomy, facilitating a more detailed analysis.

This study also has several limitations. First, the evaluation was limited to the professional medicine subset of the MMLU-Pro dataset. Although this benchmark includes complex clinical scenarios across multiple specialties, it does not fully capture the longitudinal context or multimodal information present in real-world clinical practice. The number of available questions limited the statistical power for specialty-level analyses. Qualitative error analysis only focused on a targeted subset of model outputs. Open-ended answers for one model were only reviewed by one physician reviewer. Ideally, they should be reviewed by two reviewers, with a third involved if there is disagreement for all items. In addition, benchmark-based evaluation depends on the validity of the dataset answer key. Several responses classified as incorrect by benchmark criteria were considered clinically reasonable by clinician reviewers, suggesting that some apparent errors may reflect limitations of the benchmark rather than deficiencies in model reasoning.

Future Directions

Future work should expand benchmarking of LRMs with more diverse clinical datasets, multilingual cases, and real-world clinical scenarios to better assess fairness, generalizability, and reasoning robustness [16,34]. Comparative evaluations across multiple models using standardized scenario-level protocols and blinded multi-annotator review will further clarify differences in reasoning behavior and improve methodological reliability [26,29,30]. Exploration of prompt engineering, repeated inference testing, and retrieval-augmented generation approaches may also improve model stability and diagnostic reasoning performance in clinical and educational settings [28,35].

Conclusions

LRMs demonstrated strong diagnostic performance across complex clinical scenarios in both closed-ended and open-ended question formats. These findings support their potential value as pedagogical and research tools, particularly in supervised settings. Gemini 3 Pro achieved higher overall accuracy than DeepSeek R1. These models show promise as tools for supervised medical education and research. However, further validation is needed using expanded datasets and

expert-adjudicated benchmarks. These additional steps, along with validation in real-world clinical environments, will be necessary before such systems can be considered reliable components of clinical decision support.

Acknowledgments

During the preparation of this work, the authors used GPT-4 and Perplexity AI to assist with grammatical editing, brainstorming structure, and generating R code for statistical analysis. After using these tools, the authors reviewed and edited the content as needed, verified the code and analysis, and took full responsibility for the content of the publication.

Funding

The authors received no specific funding for this work.

Data Availability

The datasets generated and analyzed during the current study are available in the GitHub repository [14]. The raw model outputs and error analysis logs are available from the corresponding author upon reasonable request.

Authors' Contributions

MB and RH conceptualized the study with input from DK and MH. RH led data curation with support from MB, DK, and MH. MB led the formal analysis and visualization, supported by RH. All authors (MB, RH, DK, MH) contributed equally to the investigation, drafting of the original manuscript, and review and editing of the final version. RH administered the project. All authors read and approved the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Final TRIPOD-LLM mapping table.
[DOCX File, 16 KB - xmed_v7i1e76822_app1.docx]

Multimedia Appendix 2
Complete table of diagnostic accuracy by clinical specialty for DeepSeek R1 and Gemini 3 Pro across closed- and open-ended formats for 6 or more cases per specialty.
[DOCX File, 18 KB - xmed_v7i1e76822_app2.docx]

References

1. Zhang K, Meng X, Yan X, et al. Revolutionizing health care: the transformative impact of large language models in medicine. J Med Internet Res 2025 Jan 7;27:e59069. [doi: 10.2196/59069] [Medline: 39773666]
2. Meng X, Yan X, Zhang K, et al. The application of large language models in medicine: a scoping review. iScience 2024 May 17;27(5):109713. [doi: 10.1016/j.isci.2024.109713] [Medline: 38746668]
3. Guo D, Yang D, Zhang H, Song J, Zhang R, Xu R, et al. DeepSeek-R1: incentivizing reasoning capability in LLMs via reinforcement learning. arXiv. Preprint posted online on Jan 28, 2025. [doi: 10.48550/arXiv.2501.12948]
4. Mercer S, Spillard S, Martin DP. Brief analysis of DeepSeek R1 and its implications for generative AI. arXiv. Preprint posted online on Feb 4, 2025. [doi: 10.48550/arXiv.2502.02523]
5. Bergmann D. What is mixture of experts? IBM Think Blog. 2024. URL: https://www.ibm.com/think/topics/mixture-of-experts [accessed 2025-04-14]
6. Zhou J, Cheng Y, He S, Chen Y, Chen H. Large language models for transforming healthcare: a perspective on DeepSeek‐R1. MedComm 2025 Jun;4(2):e70021. [doi: 10.1002/mef2.70021]
7. Cleary D. DeepSeek r-1 model overview and how it ranks against OpenAI's O1. PromptHub Blog. 2025. URL: https://www.prompthub.us/blog/deepseek-r-1-model-overview-and-how-it-ranks-against-openais-o1 [accessed 2025-04-14]
8. MMLU-Pro benchmark leaderboard. Artificial Analysis. 2024. URL: https://artificialanalysis.ai/evaluations/mmlu-pro [accessed 2026-04-10]
9. Jin D, Pan E, Oufattole N, Weng WH, Fang H, Szolovits P. What disease does this patient have? A large-scale open domain question answering dataset from medical exams. Appl Sci (Basel) 2021;11(14):6421. [doi: 10.3390/app11146421]
10. Desjardins I, Touchie C, Pugh D, Wood TJ, Humphrey-Murto S. The impact of cueing on written examinations of clinical decision making: a case study. Med Educ 2014 Mar;48(3):255-261. [doi: 10.1111/medu.12296] [Medline: 24528460]

11. Willing S, Ostapczuk M, Musch J. Do sequentially-presented answer options prevent the use of testwiseness cues on continuing medical education tests? Adv Health Sci Educ Theory Pract 2015 Mar;20(1):247-263. [doi: 10.1007/s10459-014-9528-2] [Medline: 24950895]
12. Griot M, Hemptinne C, Vanderdonckt J, Yuksel D. Large language models lack essential metacognition for reliable medical reasoning. Nat Commun 2025 Jan 14;16(1):642. [doi: 10.1038/s41467-024-55628-6] [Medline: 39809759]
13. Wang Y, Ma X, Zhang G, et al. MMLU-pro: a more robust and challenging multi-task language understanding benchmark. arXiv. Preprint posted online on Jun 3, 2024. [doi: 10.48550/arXiv.2406.01574]
14. TIGER-AI Lab. MMLU pro: code and data (version neurips 2024). GitHub. 2024. URL: https://github.com/TIGER-AI-Lab/MMLU-Pro [accessed 2025-04-14]
15. Gong EJ, Bang CS, Lee JJ, Baik GH. Knowledge-practice performance gap in clinical large language models: systematic review of 39 benchmarks. J Med Internet Res 2025 Dec 1;27:e84120. [doi: 10.2196/84120] [Medline: 41325597]
16. Wang Z, Li H, Huang D, Kim HS, Shin CW, Rahmani AM. HealthQ: unveiling questioning capabilities of LLM chains in healthcare conversations. Smart Health (2014) 2025 Jun;36:100570. [doi: 10.1016/j.smhl.2025.100570]
17. Wornow M, Bedi S, Fuentes Hernandez MA, Steinberg E, Fries JA, Re C, et al. Evaluating long context models for clinical prediction tasks on EHRs. arXiv. Preprint posted online on Dec 9, 2024. [doi: 10.48550/arXiv.2412.16178]
18. Sonoda Y, Kurokawa R, Hagiwara A, et al. Structured clinical reasoning prompt enhances LLM's diagnostic capabilities in Diagnosis Please quiz cases. Jpn J Radiol 2025 Apr;43(4):586-592. [doi: 10.1007/s11604-024-01712-2] [Medline: 39625594]
19. Gallifant J, Afshar M, Ameen S, et al. The TRIPOD-LLM reporting guideline for studies using large language models. Nat Med 2025 Jan;31(1):60-69. [doi: 10.1038/s41591-024-03425-5] [Medline: 39779929]
20. DeepSeek-R1 release. DeepSeek AI. 2025 Jan 19. URL: https://api-docs.deepseek.com/news/news250120 [accessed 2025-04-14]
21. Eriksson V. Perplexity releases a censorship-free variant of Deepseek R1. Computerworld. 2025. URL: https://www.computerworld.com/article/3829462/perplexity-releases-censorship-free-variant-of-deepseek-r1.html [accessed 2025-04-14]
22. Perplexity. URL: https://www.perplexity.ai [accessed 2025-04-14]
23. Gemini developer guide. Gemini AI. URL: https://ai.google.dev/gemini-api/docs/gemini-3 [accessed 2026-04-10]
24. McNemar Q. Note on the sampling error of the difference between correlated proportions or percentages. Psychometrika 1947 Jun;12(2):153-157. [doi: 10.1007/BF02295996] [Medline: 20254758]
25. Field A. Discovering Statistics Using IBM SPSS Statistics, 4th edition: Sage Publications; 2013. URL: https://vlb-content.vorarlberg.at/fhbscan1/330900091084.pdf [accessed 2025-09-05]
26. Moëll B, Sand Aronsson F, Akbar S. Medical reasoning in LLMs: an in-depth analysis of DeepSeek R1. Front Artif Intell 2025;8:1616145. [doi: 10.3389/frai.2025.1616145] [Medline: 40607450]
27. Xu P, Wu Y, Jin K, Chen X, He M, Shi D. DeepSeek-R1 outperforms Gemini 2.0 Pro, OpenAI O1, and O3-mini in bilingual complex ophthalmology reasoning. Adv Ophthalmol Pract Res 2025;5(3):189-195. [doi: 10.1016/j.aopr.2025.05.001] [Medline: 40678192]
28. Mondillo G, Colosimo S, Perrotta A, Frattolillo V, Masino M, et al. Comparative evaluation of advanced AI reasoning models in pediatric clinical decision support: ChatGPT O1 vs. DeepSeek-R1. medRxiv. Preprint posted online on Jan 28, 2025. [doi: 10.1101/2025.01.27.25321169]
29. Alexandrou M, Mahtani AU, Rempakos A, et al. Performance of large language models on the internal medicine mock exam. Mayo Clin Proc 2025 Mar;100(3):569-571. [doi: 10.1016/j.mayocp.2024.11.010] [Medline: 40044366]
30. Hoyt RE, Knight D, Haider M, Bajwa M. Evaluating large reasoning model performance on complex medical scenarios in the MMLU-pro benchmark. medRxiv. Preprint posted online on Apr 19, 2025. [doi: 10.1101/2025.04.07.25325385]
31. Ma W, He J, Snell C, Griggs T, Min S, Zaharia M. Reasoning models can be effective without thinking. arXiv. Preprint posted online on Apr 14, 2025. [doi: 10.48550/arXiv.2504.09858]
32. Sarkar A, Tankelevitch L, Lee H, et al. The impact of generative AI on critical thinking: self-reported reductions in cognitive effort and confidence effects from a survey of knowledge workers. Presented at: CHI '25: Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems; Apr 26 to May 1, 2025. [doi: 10.1145/3706598.3713778]
33. Bastani H, Bastani O, Sungu A, Ge H, Kabakcı Ö, Mariman R. Generative AI without guardrails can harm learning: evidence from high school mathematics. Proc Natl Acad Sci U S A 2025 Jul;122(26):e2422633122. [doi: 10.1073/pnas.2422633122] [Medline: 40560616]
34. 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]
35. Meincke L, Mollick E, Mollick L, Shapiro D. Prompting science report 1: prompt engineering is complicated and contingent. arXiv. Preprint posted online on Mar 4, 2025. [doi: 10.48550/arXiv.2503.04818]

Abbreviations

AI: artificial intelligence
LLM: large language model

LRM: large reasoning model
MCQ: multiple choice question
MedQA: Medical Question Answering
MMLU-Pro: Massive Multitask Language Understanding Pro
TRIPOD-LLM: Transparent Reporting of a Multivariable Model for Individual Prognosis or Diagnosis–Large Language Model
USMLE: United States Medical Licensing Examination

Edited by A Schwartz; submitted 01.May.2025; peer-reviewed by J You, Z Wang, M Kejriwal; revised version received 15.Mar.2026; accepted 23.Mar.2026; published 27.Apr.2026.

Please cite as:
Bajwa M, Hoyt R, Knight D, Haider M
The Performance of DeepSeek R1 and Gemini 3 in Complex Medical Scenarios: Comparative Study
JMIRx Med 2026;7:e76822
URL: https://xmed.jmir.org/2026/1/e76822
doi:10.2196/76822

Investigating the Variable Component of the Systematic Error, a Neglected Error Parameter: Theoretical Reevaluation Study

Atilla Barna Vandra, MS

Spitalul Clinic Judetean de Urgenta Brasov, Str. Berzei 2 Bl. B. ap 20, Brasov, Romania

Corresponding Author:
Atilla Barna Vandra, MS
Spitalul Clinic Judetean de Urgenta Brasov, Str. Berzei 2 Bl. B. ap 20, Brasov, Romania

Related Articles:

Companion article: https://www.medrxiv.org/content/10.1101/2023.05.24.23290382v1

Companion article: https://med.jmirx.org/2026/1/e88830

Companion article: https://med.jmirx.org/2026/1/e90221

Companion article: https://med.jmirx.org/2026/1/e88981

Abstract

Background: The existence of the variable component of the systematic error (VCSE) was known from the beginning. Still, it is a kind of taboo: it does not have a definition in the International Vocabulary of Metrology and is not present in equations, as it is considered transformed over time into random error.

Objective: This theoretical study aims to reevaluate the role and significance of the VCSE in quality control (QC).

Methods: Assuming three quintessential principles—(1) a parameter must be determined under the same conditions under which it is used, (2) a calibration cannot correct smaller biases than the calibration error, and (3) a constant cannot correct a variable—it was deduced that the source of the VCSE is bias drift caused by reagent instability and the shifts caused by human interventions. Both phenomena are mentioned in the literature. The two causes were confirmed by two series of computer simulations using 1000 normally distributed values with an SD of 1 to simulate random error and appropriately chosen bias values to simulate (1) drifts with different slopes and (2) variable shifts. Real-life examples from day-to-day QC, using Roche reagents on Cobas 6000 and Cobas PRO analyzers, confirmed the computer simulations.

Results: "The bias" is a definitional uncertainty because bias is time-variable. The causes of the cyclic variations are reagent instability and human intervention, confirmed by computer simulation and real-life QC data. Making a clear distinction between bias measured under repeatability and reproducibility within laboratory conditions, as in the case of SDs, and also separating constant and variable subcomponents of the systematic error, 2 sets of error parameters are obtained, each set being consistent with the measurement conditions. The link between them is the time-variable VCSE function. More properties of the VCSE(t) impose a distinction from random error component: predictability and corrigibility in the short term and non-Gaussian distribution. Its transformation into random phenomena is a myth based on confusion between random and variable error components. The accurate determination of the VCSE(t) function is possible, but it has an excessively high cost-effectiveness ratio. Because it is hidden in the bias measured in repeatability and in the SD in reproducibility within laboratory conditions, it helps us to avoid the redundant use in total measurement error and MU equations. Several false assumptions behind the Westgard rules were uncovered.

Conclusions: The new error model aims to serve as the foundation of a new QC system. Internal QC decisions are only consistent with graphs designed using SD measured in repeatability conditions; therefore, they are not consistent with the actual Westgard rules. Alarms should be avoided in cases of incorrigible biases. Immediately after calibration, constant biases, gradually increasing biases, and unexpected shifts in bias represent distinct situations, each requiring a unique strategy.

(JMIRx Med 2026;7:e49657) doi:10.2196/49657

KEYWORDS

repeatability condition; reproducibility within laboratory condition, measurement; systematic error; clinical laboratory; quality control; bias; QC; statistical; statistics; mathematics; computer simulation

Introduction

The author was motivated to research and publish this study after observing several statistically impossible internal quality control (IQC) graphs designed with s_{RW} (SD measured under variable conditions, reproducibility within laboratory conditions), as recommended by Westgard et al [1]. For example, there are no $R_{1\text{-}2S}$ rule violations in a month. With 180 measurements/month (Romanian laws impose 3 control runs per day), in the case of an assumed normal distribution and a correct SD, the theoretical probability (calculated using normal distribution tables) of such a graph is 0.0224%. The author observed such (and other types) of statistically impossible graphs on all analyzers he practiced: Hitachi Modular, Cobas 6000, Cobas Pro, Cobas Pure, Architect 8000, JEOL, Siemens Advia, and BTS 370.

The former statistically impossible graphs become possible if we design the quality control (QC) graphs with an overestimated SD. For example, assuming an overestimation of 50% of the SD (practically applying instead of the $R_{1\text{-}2S}$ rule, the $R_{1\text{-}3S}$ rule as a warning), the probability of no $R_{1\text{-}2S}$ rule violations in a month becomes 62.58%. The Westgard rules are only correctly applied if we design the QC graphs with the correct SD (the measure of the pure random error component [RE]).

There is no reciprocal relationship between the normal distribution and the SD. We can calculate an SD from any data set, not just from data with a normal distribution. An SD is not proof of a normal distribution. According to Stahl [2]:

> *[The name of] Normal distribution was not the luckiest choice because other distributions are perceived as abnormal.*

Consequently, scientists perceive all distributions as not abnormal and do not verify the Gaussian character. The Gauss equation is only valid if conditions do not change. Westgard rules assume a normal distribution. However, the long-term control data are not normally distributed [3,4]. The significant variation in the monthly biases and SDs also sustains the non-Gaussian distribution (see data published by Kumar and Mohan [5]).

A significant source of error is the definition of the random measurement error in the International Vocabulary of Metrology (VIM) 2.19 [6], which considers random and unpredictable terms equivalent. According to Krystek [7]:

> *We speak of 'random' variations, although we cannot explain what the attribute 'random' actually means.*

There are different types of unpredictable phenomena. Some such examples:

a. A transient phenomenon causing an outlier.
b. An unexpected phenomenon causing a systematic change (shift).
c. A cyclical (eg, sinusoidal) variation can be subjectively perceived as random if checked in more extended time frames than its period.
d. Non-Gaussian (eg, uniformly) distributed random phenomena, like the values generated by the RAND() function in EXCEL.
e. Expected change with unpredictable extent (eg, human interventions), alternating with predictable time frames. It can be named a randomly variable systematic phenomenon.
f. Typical random phenomena caused by the inconstancy of the measuring system (eg, sampling error). Only the last phenomenon is the source of normally distributed data sets.

The confusion between the typical random and the randomly variable systematic phenomena is a severe error source in the QC. The author used the following assumptions:

- Assumption 1: The systematic error component (SE) is concentration-dependent (we perform QC measurements on more levels).
- Assumption 2: The SE is time-dependent (we repeat controls periodically).
- Assumption 3: Calibration is a measurement subject to errors (after calibration, a QC run is compulsory).
- Assumption 4: The instrument is quasi-constant in time. Maintenance does not impose corrective actions (eg, recalibrations), only QC.
- Assumption 5: An instrument failure cannot cause specific systematic variations, and the errors are of aberrant size (eg, a blown lamp).

This study is consistent with the following quintessential principles valid in all sciences:

- Quintessential principle 1: We must determine all parameters under the same conditions under which we use them. For example, if we determine a parameter under specific constant conditions, we cannot use it for predictions in variable conditions. We can extend the use of a parameter obtained within a given time frame to other time frames only if we assume that it is constant.
- Quintessential principle 2: An action (eg, calibration) can efficiently correct neither smaller biases than its average error nor smaller biases than the uncertainty of the bias value.
- Quintessential principle 3: We cannot correct a variable error by adding a constant.

The SE (bias) is dependent on concentration and time (SE≈B(c, t), Assumptions 1 and 2). To apply correctly, we must modify the error model. Westgard et al [8] separated the bias into a constant component (CE) and another proportional to concentration (PE), making it possible to deal with concentration dependency. If we focus on a single control level, the separation is unnecessary. The corrected error model has a wide range of applicability [9,10].

A similar, generally accepted separation of bias components to deal with time dependency does not exist. Westgard et al [8] started from the assumption of a constant bias. As Badrick [3] observed:

> *[In the Westgard model] One assumption is that the bias is unchanged over time; 'Systematic' implies a specific point in time.*

However, JCGM −6:2020 GUM 10.6 has recommendations in the case of drift effects [11], the JCGM 100:2008 GUM 3.2.4 [12] recommendation "It is assumed that the result of a measurement has been corrected for all recognized significant systematic effects" hides a similar assumption. Neither a correction (GUM B.2.23) nor a correction factor (GUM B.2.24) can eliminate a function (a time-variable bias, quintessential principle 3). The bias is undoubtedly time-variable (Assumption 2). According to Leito [13]:

> *Bias determined within a single day is different from one determined on different days (and averaged).*

If so, the bias measured in external quality assessment (EQA) has a validity term of only 24 hours. When we obtain the result, the value is obsolete. The variable bias is neither eliminated by corrections nor by calibration because it reappears (quintessential principle 3).

When substituting the bias value into an equation, the question arises: Which bias? The bias of today, the value measured in the last EQA, or the long-term mean of the bias values? "The bias" is a definitional uncertainty that imposes a distinction between bias types and their separation into a time-invariable component (CCSE: constant component of systematic error) and a time-variable function (variable component of the systematic error=VCSE[t]). Focusing on a single control level:

(1)TE(t)=SE(t)+RE=CCSE+VCSE(t)+RE

This study aims to identify, quantify, and characterize these bias components, if possible.

The VIM 2.17 definition [14] by the "or" word indirectly defines 2 SE subcomponents:

> *The systematic measurement error is the component of the measurement error that in replicate measurements remains constant or varies in a predictable manner.*

The CCSE and the VCSE(t) are neither defined nor at least mentioned in VIM. Time variability was known from the beginning [15]. However, the phenomenon has only come into focus in recent years. Due to the lack of standardization, the authors use different names, definitions, and notations [15-23], which cause difficulties in research. The definitions are not (entirely) equivalent. Others only make the difference between short-term bias and long-term bias [9,24] or bias of the moment "t" and mean bias [25], suggesting bias variability.

Several authors built alternative error models to include the VCSE(t) function [15,19,23,25]. A particular case is the graphical model of Theodorsson et al [21], which attempts to prove that: "Variable bias components become random errors over time."

In their model, the variable bias components are included in the SE for short time frames, while in long time frames, they are included in the RE. However, the model is consistent with the VIM 2.17 definition of the SE; its accuracy is debatable because the definition does not distinguish between randomly variable systematic and typical random phenomena (cases e and f of unpredictable).

The transformation of the variable SE components into random ones is only subjective, based on an inaccurate definition. Only the long-term control data are dispersed under the influence of 2 distinct variable phenomena: the RE and bias variation (the VCSE[t]). We can calculate an SD from the VCSE(t) values, as from any variable set of data (cases b-d of unpredictable). Let us note its s_{VCSE} (the SD calculable from the daily [run] mean, bias, or VCSE[t] values). According to more authors (using different names, definitions, and notations), the link between the SD measured in repeatability and reproducibility within laboratory conditions is the s_{VCSE} [19,22,23].

(2)SRW=Sr+SVCSE=Sr2+SVCSE2

The VCSE(t) is hidden in the bias of the moment "t" and s_{RW}.

Initially, the bias variations were perceived as unpredictable. Shewhart [15] stated:

> *The causes of this variability are, in general, unknown.*

Similar opinions have been sustained by Westgard et al [8]. Recent studies identified 2 sources of bias variability. According to Marquise [16]:

> *Every new calibration creates a different bias, which appears as a random shift on the chart.*

Magnusson et al [22] referred to the phenomenon as variations in calibration over time. The consequence is an alternation between periods of constant bias with random variations in the SE.

The reagent instability causes a gradually increasing bias (in absolute values) [18,19]. The bias cannot continue to increase indefinitely because we take corrective actions. Consequently, we obtain a sawtooth-like cyclical bias variation. Mackay et al [23] acknowledge both phenomena as sources of bias and variation.

Using computer simulations and real-life QC examples, the author will analyze these phenomena in the Experimental Data section. In the Discussion section, the properties of the VCSE(t) function and the s_{VCSE} will be compared with other bias and SD components.

There are 2 points of view in the clinical laboratory. The accreditation services and clinicians are interested in the limit of credibility of the results: the measurement uncertainty. This point of view is consistent with error parameters measured in reproducibility within laboratory conditions (quintessential principle 1). Unfortunately, this point of view is imposed on all decisions, becoming a source of error.

The laboratory specialist focuses on short-term decisions: May I run patient samples now, or must I make corrective actions before? The decisions are consistent with error parameters measured in repeatability conditions, but not those obtained in long time frames (quintessential principle 1).

There are 2 conflicting approaches in the QC. Gauss [26] introduced the error approach, which was considered valid until the emergence of the measurement uncertainty (MU) approach

described by GUM [7]. Usually, there is an expectation to adhere to one of these approaches.

While the theoreticians of the uncertainty of measurement (UM) formulated some pertinent critiques, the UM theory is not perfect. The comparison of the weaknesses and strengths of the error and UM approaches is not the task of this study. Neither the UM approach can challenge the total measurement error (TE) approach-based internal QC decisions, nor can the TE approach substitute the UM in uncertainty calculations [23]. The 2 approaches link to 2 different points of view, and predictably, they will coexist as a state-of-the-art situation. The laboratory specialists must use both, depending on their tasks. Moreover, the 2 approaches share commonalities, using the same (oversimplified) error model. This study challenges the error model, influencing both approaches. The focus of this study is on short-term, internal QC decisions. Therefore, the consequences on UM calculations will only be mentioned.

Methods

This theoretical study uses mathematical statistics. Most statements and observations are present in the literature, but only as mosaic pieces. Critical statements are based on theoretical deductions, computer simulations, and observations made in the author's 40 years of experience in the clinical laboratory. Real-life examples are from the day-to-day IQC of the laboratory of the Brasov County Clinical Hospital for Urgencies (SCJUBv). The author made the exemplified measurements on Cobas 6000 and Cobas Pro analyzers using Roche reagents, but observed similar phenomena on all analyzers he worked with.

A total of 1000 data (expressed with one decimal) with normal distribution, mean 0 (SD 1), were generated to simulate RE. The bias variation was simulated by choosing bias values depending on the task. TE was calculated as the sum of the bias and RE. From the daily RE, B, and TE values, respectively, the s_r (SD measured in constant, repeatability conditions), s_{VCSE}, and s_{RW} were calculated.

To simulate the influence of a single calibration error on the SDs, the bias was maintained at 0 in the first 500 data, and the same chosen value simulating a bias was used for the last 500 in each simulation. Changing the bias from 0 to 2 (0 - 2 s_r) with increments of 0.25 (0.25s_r), 9 data sets of s_r, s_{VCSE}, and s_{RW} were obtained. The sRW2 was represented in the function of sVCSE2 (Table 1).

Table . Computer simulation of a single calibration. In each simulation, "n" takes integer values between 0 and 8 (a total of 9 values). re_i values have a normal distribution with SD=s_r=1.004.

Time (t)	RE[a]	Bias	TE[b]
1	re_1=2.1	0	2.1
2	re_2=−1	0	−1
500	re_{500}=0.1	0	0.1
501	re_{501}=1.7	n × 0.25	1.7 + 0.25 n
502	re_{502}=−0.9	n × 0.25	−0.9 + 0.25 n
1000	re_{1000}=−1.2	n × 0.25	−1.2 + 0.25 n
SD	s_r[c]=1.004	s_{VCSE}[d] = n × 0.125	s_{RW}[e]

[a]RE: random error component.

[b]TE: total measurement error.

[c]s_r: SD measured in constant, repeatability conditions.

[d]s_{VCSE}: the SD calculable from the daily (run) mean, bias, or VCSE(t) values.

[e]s_{RW}: SD measured in variable, reproducibility within laboratory conditions.

To simulate the influence of more calibration errors on the SDs, (3 random changes in the mean) were added 4 × 10 bias values (equal to 1.5, −1, −0.5, and 0) to 2 × 40 normally distributed values (real SD of 1.07), simulating RE on 2 levels. s_{VCSE} from the bias values, s_r from the RE values, and s_{RW} from the TE values were calculated in different time frames.

One thousand and one linearly decreasing bias values were chosen (from 0 to B) to simulate the influence of drift in bias. By changing the slope factor (by changing the value of Bias from 0 to 4 with increments of 0.5), 9 data sets of s_r, s_{VCSE}, and s_{RW} were obtained. The sRW2 was represented in the function of sVCSE2 (Table 2).

Table . Computer simulation of a quasilinear drift caused by reagent degradation. "b"=B/1000. In each simulation, B takes values from 0 to 4 with increments of 0.5 (total 9 values/simulations). re_i values have a normal distribution with SD=s_r≈1.

Time (t)	RE[a]	Bias	TE[b]
0	re_0=0.6	b × 0	0.6 + 0
1	re_1=−2.1	b × 1	−2.1 + b
500	re_{500}	b × 500	re_{500} + b × 500
999	re_{999}=−0.8	b × 999	−0.8 + b × 999
1000	re_{1000}=−1.2	b × 1000	−1.2 + b × 1000
SD	s_r[c]=1.004	s_{VCSE}[d]	s_{RW}[e]

[a]RE: random error component.

[b]TE: total measurement error.

[c]s_r: SD measured in constant, repeatability conditions.

[d]s_{VCSE}: the SD calculable from the daily (run) mean, bias, or VCSE(t) values.

[e]s_{RW}: SD measured in variable, reproducibility within laboratory conditions.

In the real-life data example with drift, the run mean was estimated with the SLOPE and INTERCEPT functions in Excel. A single estimated mean was calculated from the average of the run results expressed as a percentage. The CV_r (CV measured in constant, repeatability conditions) values for each level were calculated from the deviations from the estimated run mean.

The average CV_r for the whole period was calculated as the SD of the half differences of the percent expressed results (an adaptation of a method described in Nordtest 537 TR [22]).

Results

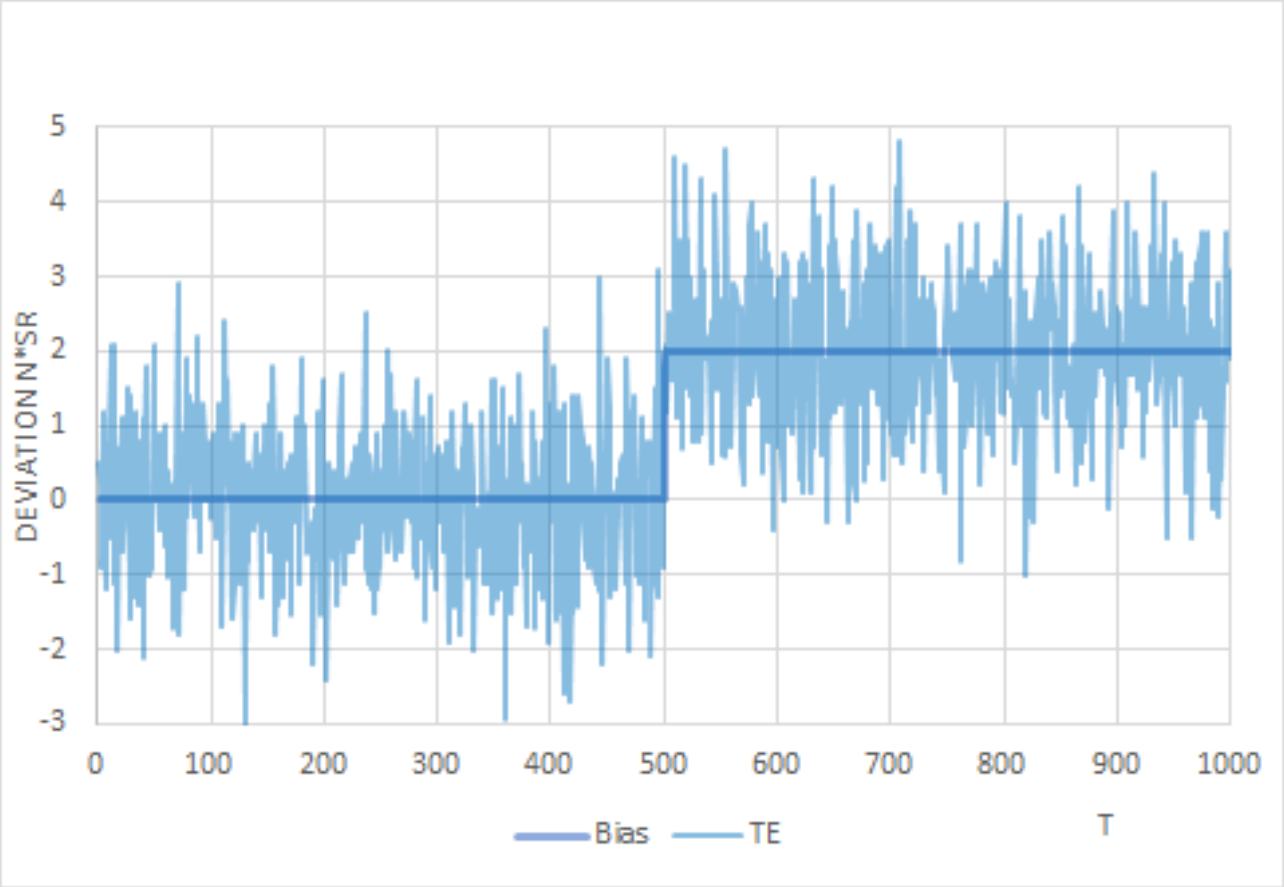
Overview

The computer simulations aimed to demonstrate that the sources of bias variation described in the literature are the true causes of the increased SD in more extended time frames and to confirm the validity of Equation 2. The real-life QC examples demonstrate that computer simulations are grounded in reality.

The Influence of a Single Shift in the Mean Caused by a Calibration

In the computer simulation of a single calibration (a single shift in the mean), the graph of the run mean is a horizontal line with bias=0 before the mean shift (calibration) and a horizontal line with mean=bias after the mean shift (calibration). The results are randomly dispersed around the run mean with SD of 1 (=s_r) (Figure 1). The SDs calculated from 500 data before and from 500 data after the shift are 1 (=s_r), while the SD calculated from all data (s_{RW}=1.43) is significantly bigger according ($F_{0.95, 500,500}$=1.43). The SD calculated from runs 480 - 520 (including the shift) is 1.55, suggesting that the bias variation causes the increase of the SD (s_{RW}). A sudden change of 1 SD (1s_r) in the mean causes an increase of only 12% in the overall SD (s_{RW}), and it is difficult to observe visually such minimal increases.

Figure 1. Computer simulation: a shift in the mean causes an increase in the s_{RW} (SD measured in variable, reproducibility within laboratory conditions). Bias variation=$2s_r$ case. TE: total measurement error.



Representing the sRW2 values as a function of sVCSE2, a linear graph with slope ≈1, consistent with Equation 2, was obtained, confirming its validity (Figure 2).

An example of magnesium obtained in March 2021 on a Cobas Pro analyzer (2 × 7 runs, 2 levels, and one calibration after 7 runs) is presented in Figure 3 to exemplify real-life data. The results were represented in %, not as absolute values, to reduce the influence of the s_{RW} variability.

Figure 2. Variation of square s_{RW} as a function of square s_{VCSE}. The slope is 1. s_{RW}: SD measured in variable, reproducibility within laboratory conditions; s_{VCSE}: SD calculable from the daily (run) mean, bias, or VCSE(t) values.

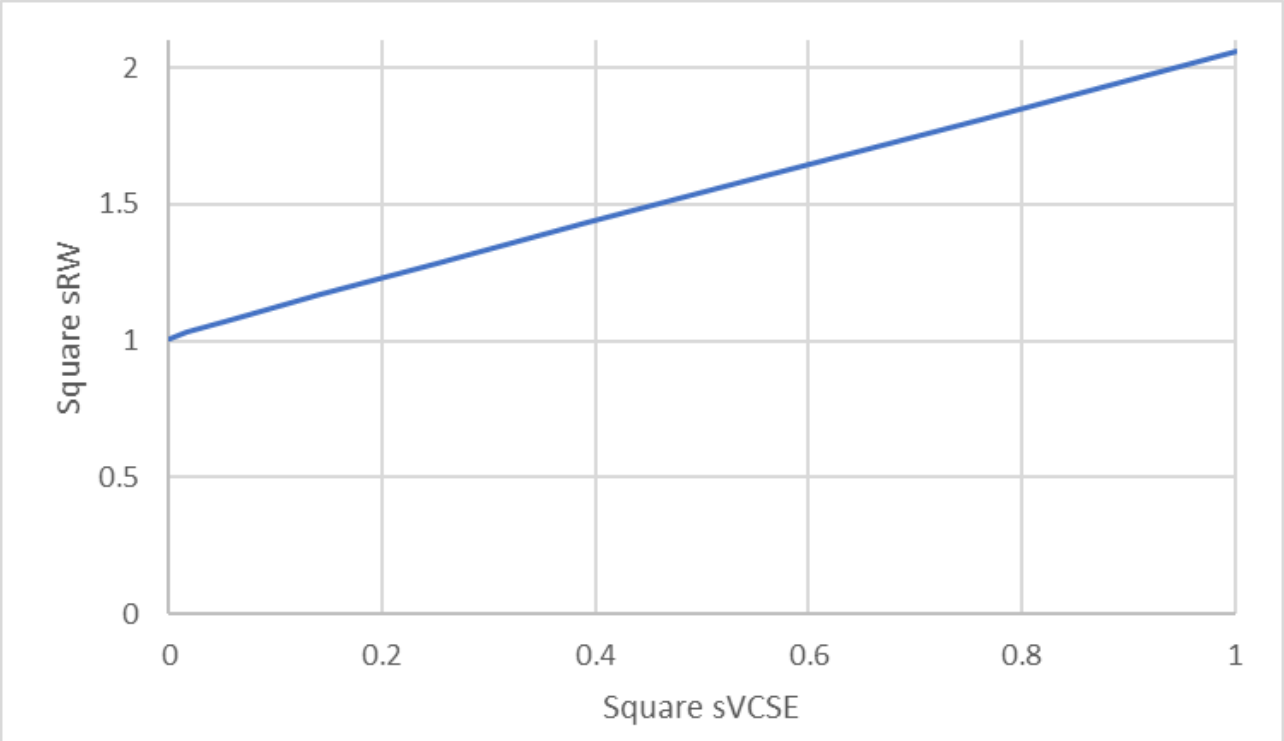
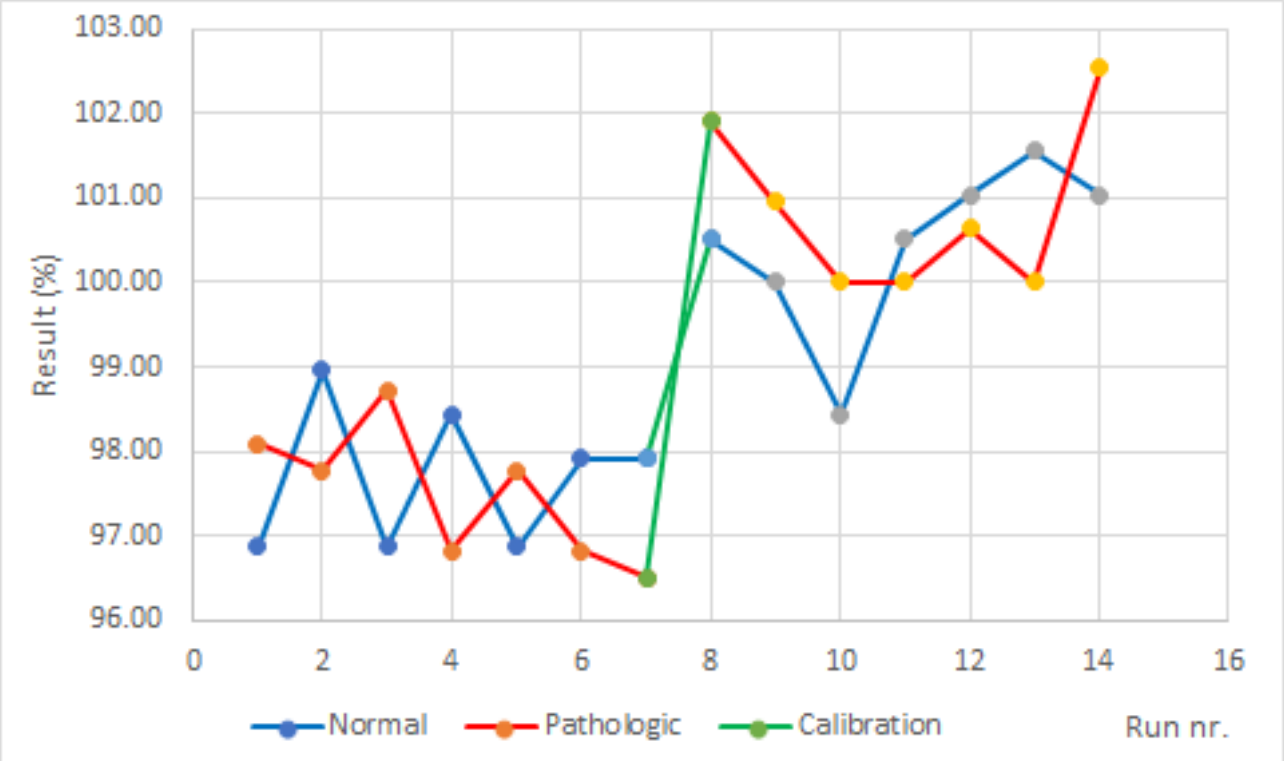


Figure 3. Calibration parameter changes cause bias variations (VCSE(t)). Real-life data. VSCE(t): variable component of the systematic error, a time-variable function.



The graph has an insignificant drift on both levels. Calculations presented in Table 3 show that before and after calibration, the coefficient of variation (CV) is consistent with the CV_r (an *F* test did not reveal significant differences), and the increase in the CV_{RW} (CV measured in variable, reproducibility within laboratory conditions) is due to the shift in the mean. Equation 2 is valid. From the s_{VCSE} calculated from the mean variation and the s_r values, it was possible to predict the value of the CV_{RW}. The *F* test did not find significant differences between the CV of all data, the predicted CV (Equation 2), and the actual CV_{RW}. The actual CV_{RW} (determined from one month's data)

is slightly bigger because it includes more calibrations and reagent changes.

Table . The increase of the s_{RW}/CV_{RW} (SD measured in variable, reproducibility within laboratory conditions/ coefficient of variation measured in variable, reproducibility within laboratory conditions) caused by a shift in the mean (calibration) can be predicted by Equation 2 (real-life data, magnesium, Cobas PRO).

Analyte and data	Number of data	CV (CV_r)[a], %	CV_r (method validation), %	CV_{VCSE}[b] =ΔB%2, %	CV of all data (CV_{RW}), %	Predicted CV_{RW} (Equation 2), %	Actual CV_{RW}, %
Mg level 1							
Before calibration	7	0.86	—[c]	—	—	—	—
After calibration	7	1.01	—	—	—	—	—
All data	14	0.94	1.24	1.39	1.69	1.95	2.14
Mg level 2							
Before calibration	7	0.83	—	—	—	—	—
After calibration	7	1.01	—	—	—	—	—
All data	14	0.92	1.11%	1.69	1.95	2.18	2.54

[a]CV_r: CV measured in constant, repeatability conditions.

[b]CV_{VCSE}: CV of the VCSE(t), s_{VCSE}, expressed as a percent of the target value.

[c]Not applicable.

The Influence of More Random Changes in the Mean (More Calibrations)

Figure 4 shows the simulation graph of more random changes in the mean. Without computer assistance, we can visually detect only the significant mean variation (run 11). As shown in Table 4, the s_r values are quasi-constant. Simultaneously, the s_{RW} values depend on the time frame (variations from 1.10 to 1.94). The bigger the mean change, the bigger the s_{RW}. The validity of Equation 2 is maintained (compare line 4 with line 10).

Figure 4. The influence of multiple mean changes (computer simulation); only significant shifts can be visually observed (run 10 - 11), and not those that are less significant. s_r: SD measured in constant, repeatability conditions.

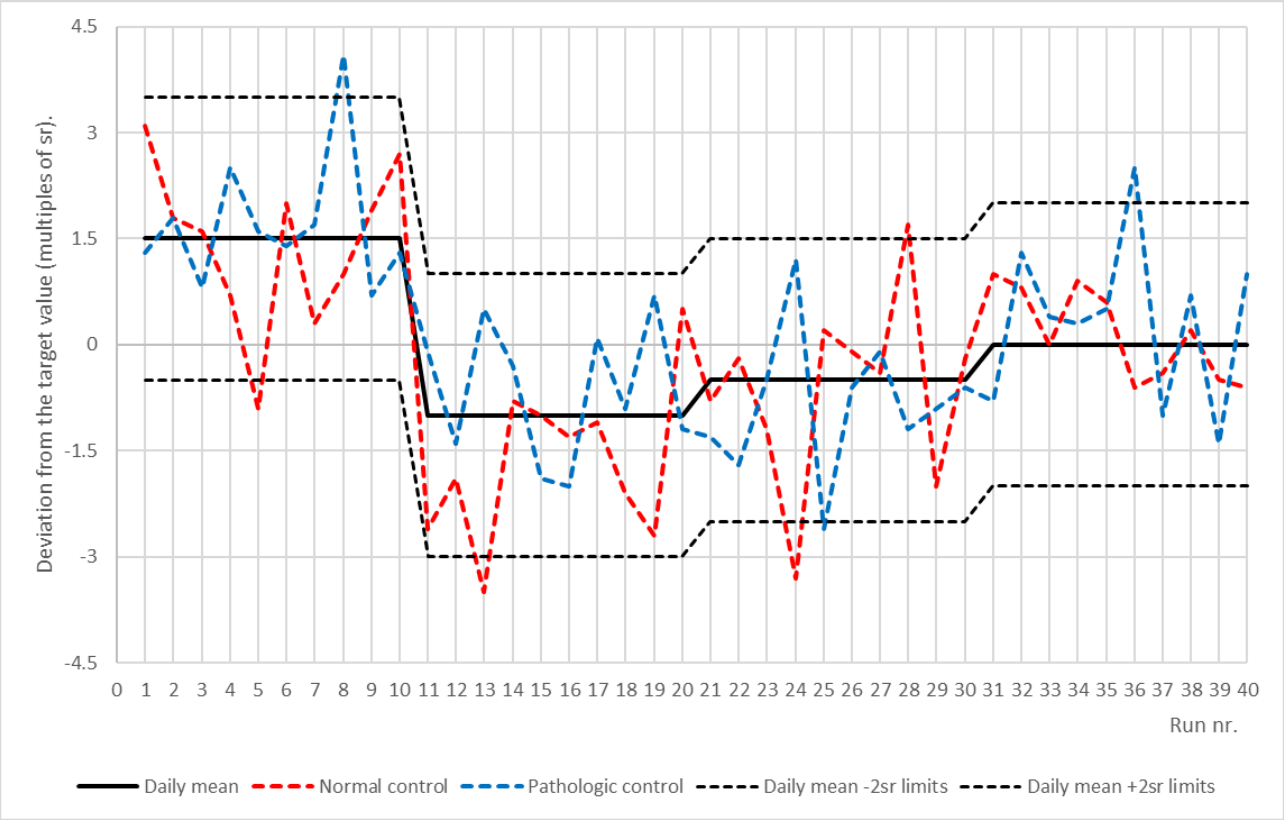


Table . There are significant differences in the s_{RW} (SD measured in variable, reproducibility within laboratory conditions) values, depending on the time frame, while s_r (SD measured in constant, repeatability conditions) in the limits of the statistical methods remains constant.

Variable and runs (time frame)	Normal	Pathologic
s_{RW}		
1 - 20	1.94	1.54
21 - 40	1.10	1.22
11 - 30	1.32	0.97
s_{RW} all		
1 - 40	1.56	1.43
s_r		
1 - 40	1.10	1.03
1 - 20	1.17	0.95
11 - 30	1.15	1.02
21 - 40	1.03	1.12
S_{VCSE}[a] (SD of bias variation)		
1 - 40	0.95	0.95
s_{RW} calculated/predicted (Equation 2)		
1 - 40	1.45	1.40

[a]s_{VCSE}: SD calculable from the daily (run) mean, bias, or VCSE(t) values.

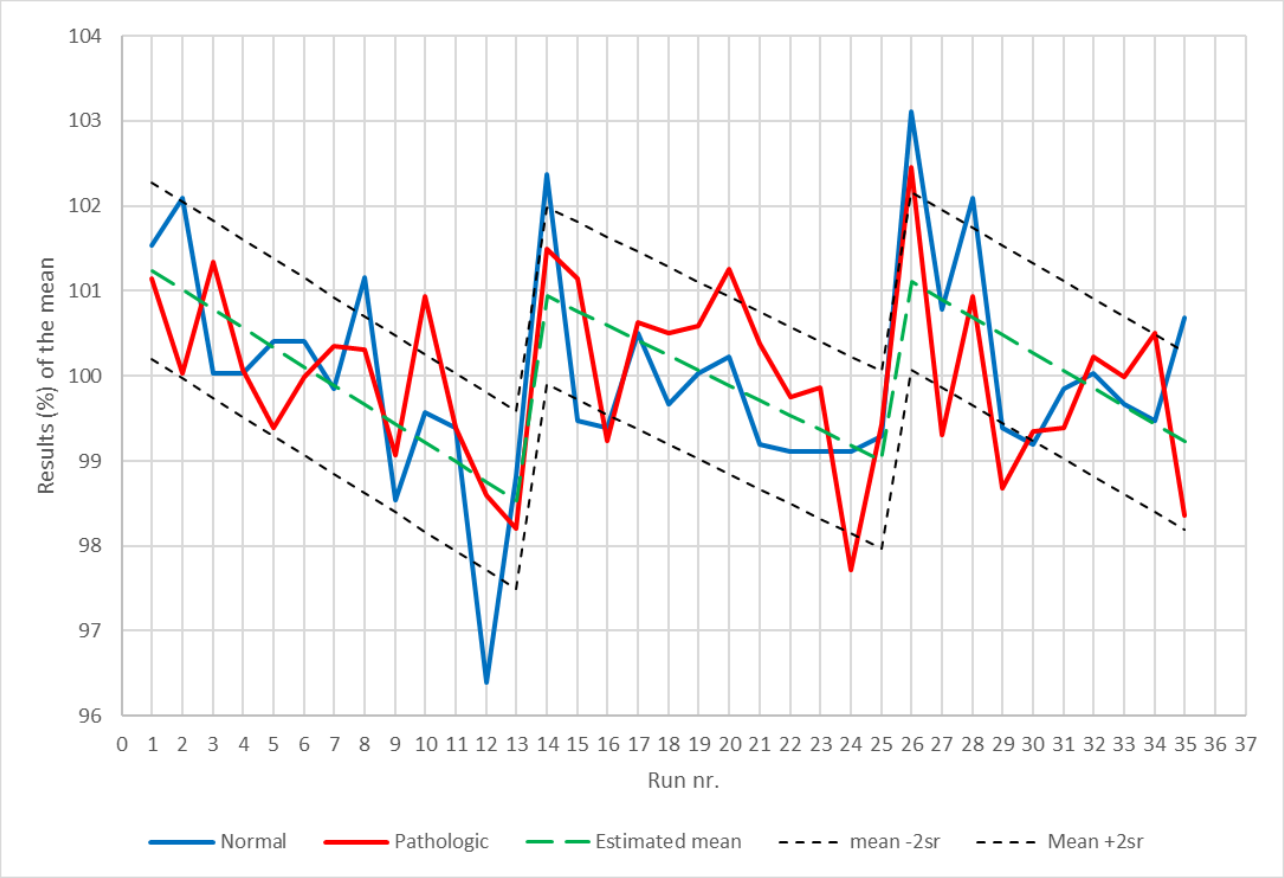
The Influence of Gradual Mean Changes (Drifts) Caused by Reagent Degradation

In the computer simulation, the graph of the daily mean was an oblique line with decreasing tendency, with slope=−0.001 × $_{max}$Bias. $_{max}$Bias is the maximum bias in absolute values in each simulation. The SD calculated from the daily means was s_{VCSE}=Biasmax2*3, corresponding to a uniform distribution. The deviation of the results from the daily means had an SD ≈1 (=1s_r) in all simulations. The SD calculated from all 1001 data (s_{RW}) was bigger than 1s_r. A bias variation of 1.5s_r caused an increase in s_{RW} of only 10%, which was difficult to observe visually.

If we represent the sRW2 values as a function of sVCSE2, we obtain an identical graph, as shown in Figure 2, consistent with Equation 2 (a linear graph with slope ≈1 and intercept ≈1).

Figure 5 shows a 35-run real-life chart (glucose, Cobas 6000 analyzer, July 2023). The period includes 2 reagent changes (corresponding to the shifts in the mean between runs 14 - 15 and 25 - 26). No calibrations were made. In the periods between reagent changes, the means are similar in all time frames (0.22%/run, 0.23%/run, and 0.20%/run), consistent with the degradation tendency of the reagent. Most data are within the estimated mean (SD 2CV_r) limits, suggesting that CV_r (s_r) is the true measure of the RE.

Figure 5. Real-life data sustain the influence of the mean drift on the variable component of the systematic error. s_r: SD measured in constant, repeatability conditions.



The CV_r values calculated from the deviations of the percent expressed results from the estimated mean are similar to the CV_r value calculated from the half differences between the percent expressed results obtained on the 2 control levels (Table 5; a Cochran *F* test for equality of 2 variances did not find significant differences between the CV_r values).

Table . The coefficient of variations (CVs) calculated from the deviations from the estimated means are similar to CV_r (CV measured in constant, repeatability conditions; in the limits of the statistical methods; CV_r [half difference, all runs]=0.73%). The CV_{RW} (CV measured in variable, reproducibility within laboratory conditions) is significantly bigger.

	Runs (%)				CV_r (method validation) (%)	CV_{RW} (%)
	1 - 13	14 - 25	26 - 35	All runs		
Normal	0.96	0.72	1.04	0.90	0.81	1.24
Pathologic	0.73	0.80	1.07	0.86	0.80	1.10

A control material handling error (reused control material) in run 26 (false simultaneous increase) caused the slightly bigger s_r in runs 26 - 35.

Another example with total bilirubin was published by Vandra [27] in a preprint paper.

Discussion

Principal Findings

"The bias" is a definitional uncertainty. The same distinction is necessary between the biases obtained in repeatability and respective reproducibility within laboratory conditions, as in the case of SDs. The need for standardization imposes similar notations. We must highlight the time-variable function character of the bias as well. The author proposes the following notations:

- $B_r(t)$=Bias measured in repeatability conditions, at the moment t.
- B-RW=Mean bias measured in reproducibility within laboratory conditions. It is the mean of the $B_r(t)$ values in a given time frame. An accent highlights the fact that it is a mean.

We can obtain only a mean bias value in more extended time frames.

A Corrected Error Model

The difference between $B_r(t)$ and B-RW is VCSE(t), a time-variable function.

$$(3)\quad VCSE(t)=Br(t)-\bar{B}RW=(Br(t)-CCSE)$$

Variations in the mean caused by reagent property changes cause drifts. The VCSE(t) cannot increase indefinitely (in absolute values) due to human interventions. It may have only cyclical variations. The cycles depend on external factors (eg, the rhythm of reagent use, frequency of human interventions, and the size of random calibration errors). They have different amplitudes, means, and lengths.

In some cases, a cycle may last even a month. The graphs of the daily means (not of the results) have sawtooth shapes masked by the noise of the RE (easily observed in the case of the unstable reagents, eg, Figure 5). In short or medium time frames, the B-RW values may have variations. The longer the time frame, the less uncertainty there is for the B-RW values. Only yearly B-RW values can be considered quasi-constant [21] and used for accurate corrections. In a chosen time frame, we can identify B-RW with the CCSE. Consequently, the mean of the VCSE(t) is 0. If we calculate the long-term mean of the $B_r(t)$ values:

$$(4)\quad \sum_{t=1}^{n} B_r(t) = \sum_{t=1}^{n} (\bar{B}RW + VCSE(t)) = n\bar{B}RW + \sum_{t=1}^{n} VCSE(t) = \bar{B}RW = CCSE = TE\bar{R}W$$

We obtain the same value for the long-term mean of TE (TE-RW)because $\sum_{t=1}^{n} RE(t)=0$. Similarly, the SD can be calculated from long-term data (s_{RW}):

$$(5)\quad s_{RW}^2 = \frac{\sum_{t=1}^{n} (B_r(t) - \bar{B}RW)^2}{n-1} = \frac{\sum_{t=1}^{n} (VCSE(t))^2}{n-1} = s_{VCSE}^2$$

Which confirms the validity of Equation 2 (because the long-term mean of RE and VCSE(t) is 0, $\sum_{t=1}^{n}(RE(t)*VCSE(t))\approx 0$). Regrouping the terms in Equation 5 can be calculated using s_{VCSE}.

Regrouping Equation 3 and adding RE to both parts of the equation yields:

$$(6)\quad TE(t)=CCSE+VCSE(t)+RE(t)=\bar{B}RW+VCSE(t)+RE(t)$$

Equations 3, 5, and 6 define a new error model, which is presented in Figure 6.

Figure 6. A new error model, taking into account the time variability of the bias. $B_r(t)$: bias measured in repeatability conditions at the moment t (a time-variable function); B_{RW}: long-term mean bias, measured in RW conditions, a constant; CCSE: constant component of systematic error; SE: systematic error component; s_r: SD measured in constant, repeatability conditions; s_{RW}: SD measured in variable, reproducibility within laboratory conditions; s_{VCSE}: SD calculable from the daily (run) mean, bias, or VCSE(t) values; TE: total measurement error; VCSE: variable component of the systematic error.

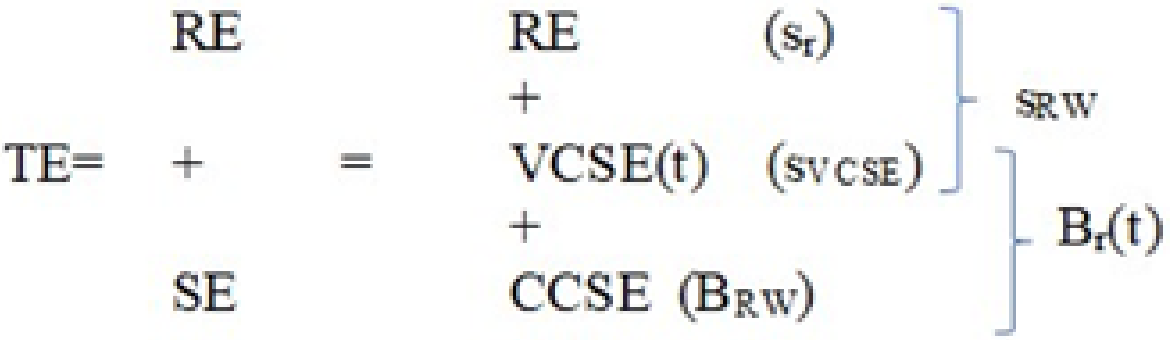


Figure 6 shows that both s_{RW} and $B_r(t)$ include VCSE(t) in a hidden form.

Two Points of View, Two Sets of Error Parameters

We obtain 2 sets of error parameters by separating the bias into a constant and a variable component and distinguishing bias measured in repeatability and reproducibility within laboratory conditions. According to quintessential principle 1, UM

calculations must be based on parameters determined in reproducibility within laboratory conditions (s_{RW}, B-RW). In the meantime, the internal QC decisions must be based on parameters determined in repeatability conditions (s_r, $B_r(t)$). The second conclusion contradicts the recommendations of Westgard et al [1] to design Levey-Jennings charts with an SD calculated from long-term control data (s_{RW}).

Proposed Definitions of CCSE and VCSE(t)

Consistent with the VIM 2.17 definitions [14], we can define the bias components as:

> *The constant component of SE (CCSE) is the component of measurement error that in replicate measurements remains constant.*

Note 1: The CCSE is the long-term mean bias B-RW, depending on the time frame.

> *The variable component of SE (VCSE(t)) is the component of measurement error that in replicate measurements varies predictably.*

Note 2: VCSE(t) is a time-variable function.

Note 3: VCSE(t) is hidden in $B_r(t)$ and s_{RW}.

The simultaneous use of $B_r(t)$ and s_{RW} causes a redundant use of VCSE(t) in equations—for example, ${}_{max}TE=B_{EQA} + z \times s_{RW}$. B_{EQA} is the bias measured in the last EQA round in repeatability conditions, and ${}_{max}TE$ is the TE limit, which includes all TE values with confidence corresponding to z, the confidence factor.

If bias is variable, TE is also variable (contradicting the graphical model of Theodorsson et al [21]). A distinction is necessary between:

a. TE of a given measurement ($TE(t)=B_r(t)+RE$). It has no practical value.
b. The maximum TE value at the moment *t* measured under repeatability conditions with a chosen confidence level.

$$(7) \max TE(t)=Br(t)+z*sr$$

Internal QC decisions must be based on ${}_{max}TE(t)$, the maximum value of the TE at the moment t of decisions with a chosen confidence level, where z is the confidence factor.

a. The maximum TE value in long time frames is measured in reproducibility within laboratory conditions with a chosen confidence.

$$(8) \max TERW=\bar{B}RW+z*sRW$$

Where ${}_{max}TE_{RW}$ is the maximum TE value in long time frames, with a chosen confidence. It must be used when setting limits and is a starting point for uncertainty of measurement (UM) calculations.

TE is also an ambiguous term. It is necessary to specify which TE is mentioned.

TE was the dominant paradigm until the emergence of UM after the publication of GUM in 1993 [11,12,28].

UM mathematically expresses our lack of knowledge about the accuracy of the result. According to VIM 2.26 [14]:

> *Uncertainty of measurement is a non-negative parameter characterizing the dispersion of the quantity values being attributed to a measurand, based on the information used.*

The definition is also mentioned in ISO 15189. According to ISO 15189, 5.6.2 [29]:

> *Sources that contribute to uncertainty may include sampling, sample preparation, sample portion selection, calibrators, reference materials, input quantities, equipment used, environmental conditions, condition of the sample and changes of operator.*

Surprisingly, neither the calibration error nor the reagent instability is mentioned among the uncertainty sources. According to the Hong Kong Association of Medical Laboratories, "the IQC procedure is designed to detect variations in reagents or calibrators" [30].

According to Magnusson and Ellison [28]:

> *The principles laid down by GUM are recognized to apply to all types of quantitative measurements, in all fields of application, and are widely accepted.*

A prerequisite for the application of the GUM [11] is that:

> *The result of a measurement has been corrected for all recognized significant systematic effects.* [GUM 3.2.4]

Using either a correction (GUM B.2.23) or a correction factor (GUM B.2.24) [11]. Then, the uncertainty of the correction is included in the uncertainty budget. Unfortunately, according to Magnusson and Ellison [28]:

> *…instances in which bias is known or suspected, but in which a specific correction cannot be justified, are comparatively common. The ISO Guide to the Expression of Uncertainty in Measurement does not provide well for this situation.*

The uncorrected bias must be included in the uncertainty budget, and due to the VCSE, it is not negligible. There is a debate in the literature about incorporating the uncorrected bias in the expression of total uncertainty (eg, Magnusson and Ellison [28], Westgard [31]). A review of this debate is not the task of this study.

UM equations start from the same error model, and TE equations ($TE=SE+RE$). They substitute the error parameters with the uncertainties caused in patient results.

$$(9) UM=UtotSE+UtotRE$$

Where $Utot_{SE}$ is the total uncertainty of the patient's result caused by the SE, and $Utot_{RE}$ is the total uncertainty caused by the RE.

There are 2 types of uncertainty in the case of both parameters: the uncertainty of the result because the error parameters exist, and our uncertainty about the value of the parameters. For example, the uncertainty of a patient's result, caused by the RE, is as follows:

$$(10) URE=z*SD$$

Unfortunately, the SD value is not accurate. Therefore:

(11)UtotRE=URE+USD=z*SDmax

where U_{SD} is the uncertainty of the SD value, and SD_{max} is the maximum value of the SD. However, the adepts of the UM critique TE theory because TE equations do not include the uncertainty of the error parameters, nor do UM equations include the uncertainty of the SD. However, s_{RW} has big monthly variations [5].

The uncertainty caused by the SE (bias) equals the bias value. U_{SE}=B. Because the bias value is uncertain, it must be added to the U_B term.

(12)UtotSE=USE+UB=B+UB

According to the GUM recommendations, all discovered bias sources must be corrected. The bias becomes insignificant, and the B term can be neglected.

Applying the UM, the first step is to correct for bias (if possible and recommended). Having 2 sets of error parameters, according to the presented error model and 2 TE equations, 2 different UM equations can be obtained.

The first, calculated in repeatability conditions, starts from Equation 7. The bias, which must be corrected in the first step, is the average of the bias measurements in the same EQA round (B-EQA). The bias value after correction can be considered negligible. The uncertainty of the correction can be determined in 2 ways (bottom-up and top-down methods). In the bottom-up approach, the bias uncertainty is calculated as the sum of uncertainties of the reference value and the uncertainty of the measurement in repeatability conditions (s_r); in the top-down method, as the sum of the uncertainty of the reference value and the root mean square of the corrected bias values ((BEQAi -B-EQA) (Where BEQAi are the individual bias results.) The 2 methods give similar results (within the statistical methods' limits) because RMS(BEQAi -B-EQA)≈s_r.

(13)UM(EQA)=USE+URE=B+z(uCref2+urec2+s2rmax)/n≈B+z(uCref2+urec2+RMS2(BEQAi−B̄EQA))/n

Where n is the number of measurements, srmax is the estimated maximum value of the s_r, uCref is the uncertainty of the nominal value of the reference material, and urec is the uncertainty of its reconstitution, equal to the uncertainty of 2 volume measurements (≈2 × 0.5% — the accuracy of the actual pipettes is 0.5% ‐ 0.6%). However, u_{rec} is not a negligible value; the recommended uncertainty equations do not include it. The division of the uncertainty of the bias with n was necessary because the bias value is a mean. As the number of measurements increases, the uncertainty of a mean value decreases n times). An equivalent equation was published by White [32] and in Nordtest TR 537 [22], except for the neglected u_{rec} value.

In repeatability conditions, the bottom-up and top-down methods, within the limits of the statistical measurements, give similar results for the uncertainty because RMS(BEQAi-B-EQA)2n≈ sr2n. The SD is an RMS of the deviations from the mean, with a correction: n is substituted with n-1. If a calibration is made between measurements, the top-down uncertainty will be bigger due to the bias variability. This is similar to the case of Mg: Table 3 and Figure 3.

Unfortunately, Equation 13 has no practical value in the clinical laboratory. There is a significant delay between the measurement and the moment when the results are obtained. In the meantime, reagent changes and calibrations are done, and the bias is changed. A constant cannot correct a variable. In addition, there is insufficient information to determine whether the bias is constant or proportional. Due to bias variability, the calculated uncertainty value cannot be used for extended time frames. UM is a long-term parameter.

The situation changes over time. The B-RW=CCSE is a constant, which can be corrected without contradicting quintessential principle 3.

Each EQA round measures a different bias using different reference materials with different u_{Cref} and with varying errors of reconstitution. The average of the measured bias values in different rounds is B-RW (absolute mean bias). Starting from Equation 8, with bottom-up and top-down approaches, we obtain:

(14)UM(tot)=USE+URE=B̄RW+z(RMSuCref2+s2RWmax)=B̄RW+z(RMS2(BEQAi−B̄EQA)+s2RWmax)

Actual recommendations suggest calculating the uncertainty of the bias correction as the root mean square (RMS) of the bias values [22], but this equation assumes "…a variance of bias based on assumed mean of zero" [28].

The assumption is only valid, and the equation is correct if the bias is corrected efficiently. If not, RMS_{bias} is not only u_B but includes the mean bias in its expression.

(15)RMSB=B̄RW2+uB2=(B̄RW2+RMS2(BEQAi−B̄EQA))

Which is only correct if we accept the quadratic addition law between bias and its uncertainty (questioned by the debates in the literature).

If n=1 and the B-RW term is added quadratically to the other terms under the square root, the top-down term of Equation 14 is equivalent to the equation proposed by Nordtest TR 537, except for the missing RMSuCref2term. The equation in Nordtest TR 537 expresses the uncertainty of a single value, not the uncertainty of a mean (n=1) [22].

(16)Ubias(true)=RMSB2+RMSuCref2=B̄RW2+RMS(BEQAi−B̄EQA)2+RMSuCref2

In repeatability conditions, the u_{Cref} and u_{rec} caused an unknown bias in the bias value, and these terms expressed our uncertainty about this value. Making more measurements decreases the influence of random errors; however, our uncertainty about the reference value remains unchanged. In the case of different EQA rounds, these biases of the bias values are variable and contribute to the bias variability. Therefore, to avoid redundancy, the RMSuCref2 term (included in RMS_B) must be eliminated from the top-down equation. While uCreforRMSuCref,) are bottom-up parameters, whereas the RMS_B is a top-down parameter, considering the consequences of the individual sources. Their mix causes redundancy in equations.

While the uncertainty caused by the bias variability (the $s_{VCSE)}$ term) is included in both expressions in the sRWmax, the top-down values are significantly bigger than the bottom-up ones. In the meantime, in the case of calculations based on

internal QC data, there are no significant differences (as in the case of EQA in a single round).

Table 6 presents the differences between the bias uncertainty results obtained with top-down and bottom-up methods. Similar calculations based on internal QC data and those from a single EQA round are provided for comparison. The number of measurements is considered n=1 in all calculations for the sake of better comparison.

Table . Differences between the uncertainty results on 2 analyzers and 5 analytes obtained in different conditions (real-life data). All values are in percentages.

Conditions/analyte	Cobas 501 13 EQA[a] rounds top-down	Biomajesty 13 EQA rounds top-down	Cobas 501 Bottom-up	Biomajesty Bottom-up	Cobas 501 Internal QC[b] bottom-up	Cobas 501 1 EQA round repeatability
ALT[c]	4.1	5.4	2.38	4.2	2.27	1.80
AST[d]	3.07	4.3	1.91	2.04	1.73	1.62
Glucose	2.1	4.02	1.94	1.9	1.71	1.26
Urea	2.8	5.59	2.5	2.04	2.39	1.38
Potassium	1.66	1.37	1.66	1.36	1.34	1.12

[a]EQA: external quality assessment.

[b]QC: quality control.

[c]ALT: alanine aminotransferase.

[d]AST: aspartate aminotransferase.

In long time frames (more EQA rounds), the uncertainty is more significant than in a single round because variable bias values are measured. The differences between internal QC and the bottom-up method are not significant and are caused by the u_{Cref} and u_{rec} included in the bottom-up uncertainty. Except for potassium, in almost all cases, the top-down method gives a bigger value due to the difference between the declared and true u_{Cref} values.

In the bottom-up equation, the declared uCrefvalue is substituted. In the meantime, the top-down equation includes the real one in the RMS_B term, causing the differences. There are 2 conditions for a correct EQA. The sample must be commutable and must have predetermined values [33]. Neither of these conditions is fulfilled in EQA with surrogate reference values (the mean of participants). The equation used to evaluate the uncertainty of the reference value may only be correct if the peer groups are homogeneous, and they are not [34]. This error causes an additional and significant uncertainty.

The uncertainty equations can be corrected by eliminating the confusion hidden in the bias definitional uncertainty. A key conclusion: only the long-term mean biases can be corrected efficiently. Correcting individual values is risky due to the variability of bias and delay. The actual EQA bias determinations conceal a significant source of uncertainty: the uncertainty of the surrogate reference values. Not even the bias variability can explain the differences between the uncertainties calculated in single and multiple EQA rounds, as well as between bottom-up and top-down methods. Following studies are necessary to sustain the former theoretical conclusions; the proofs and discussion do not fit within the limits of this study.

The existence of the VCSE suggests a change in the point of view. Even after correction, the bias reappears due to its variable properties. The confusion between the bias and the mean of the variable bias is a source of error.

The (immediately) incorrigible biases bring to attention the debates about including uncorrected biases in uncertainty equations. If they are not corrected immediately, the mean bias must be included in the uncertainty budget.

Sources of Bias Variations

We cannot quantify the preanalytical and postanalytical errors in the QC, nor can we measure the method and matrix errors only in EQA. The analytical errors detectable in IQC are:

- Environmental errors
- Laboratory errors
- Human (operator) errors
- Noninstrumental errors
- Instrumental errors [21,24]
- Rounding errors

In the case of a laboratory with air conditioning, using liquid phase reactions in thermostated conditions, the influence of the environment is quasi-negligible. The laboratory and human errors are redundant in the list. Neither specific laboratory nor specific human errors exist. Laboratory and human errors are a sum of preanalytical, noninstrumental, and instrumental errors.

We can include rounding errors in the instrumental error category. They have similar properties (both are nonspecific and time-invariable).

The instrumental errors are linked to the construction and functionality of the analyzer. They are always constant and nonspecific (assumptions 4 and 5). An instrumental failure will influence all measurements in an aberrant manner. Instrumental errors may be the sources of the constant error components, but never of the variable ones.

There are only 2 noninstrumental error sources: the reagent stability and the calibration graph (see quote from HKALM recommendations [30]). Both are specific and variable. Each

measurement has its specific reagents with variable properties. Producers only guarantee that we can successfully recalibrate the reagents in the validity term, not that the properties remain constant. Random changes in the reagent properties contradict the laws of chemistry. The changes are always unidirectional and gradual. The variation is not perfectly linear; however, linearity is an acceptable approximation in short intervals. The phenomenon is consistent with the linear bias variation model of J. Krouwer ($B=B_0+b_1t$) [19]. It applies only to time frames that do not include human interventions (such as calibrations, reagent changes, or control bottle changes).

The noise of the RE usually covers the drift. We can observe only significant drifts (if the mean change is $>1.5s_r$); however, all contribute to the increase of the s_{RW}. The significant drifts cause R_{7T}, $R_{2\text{-}2S}$, and $R_{1\text{-}3S}$ violations.

Many authors consider the calibration a quasi-perfect process [35]. Raúl Girardi, on an IFCC webinar (Metrology and uncertainty, August 21, 2021), even presented an alternative equation that reduced the bias uncertainty to the nominal value uncertainty of the reference material. Other authors share similar opinions [29]. Such an attitude neglects the most significant causes of the calibration graph error. On one hand, the measured reference material does not have the same composition as the material analyzed by the producer. It undergoes a lengthy process before being measured. Even if we neglect human errors (stability, homogenization, temperature errors), the reconstitution includes 2 volume measurements: one at the producer and another at the user. Badrick [36], referring to the Tietz Textbook of Clinical Chemistry [37], underlines:

> *The act of reconstitution can introduce an error far greater than the inherent error of the rest of the analytical process.*

Each reconstituted reference material bottle has a different concentration. We generate similar systematic errors until we use the same reconstituted calibrator bottle.

On the other hand, calibration is a measurement subject to systematic and random errors. In a linear calibration, we make 2 × 2 measurements and calculate the slope factor as a difference. We make calibrations in repeatability conditions. The average calibration random error is ≈2∗12CVr=1CVr (the error of the null-point absorption A_0 was neglected in the former estimation).

The calibration error introduces a systematic error in measurements, which remains constant within the time frame between calibrations, but each calibration induces an unpredictable variation in the systematic error. The result is a randomly variable systematic error. The phenomenon is consistent with the models presented by Marquise [16] and Magnusson et al [22]. We can observe only the significant shifts in the mean (>1 sr).

Because we can observe only the significant drifts and shifts, we tend to consider the bias variations unpredictable (case b of unpredictable), contradicting the bias definition (predictable).

The mostly predictable character of the bias suggests that a focus change in the internal QC is necessary. The QC system must also have a strategy to predict bias variations and detect unpredictable changes.

Properties of the VCSE(t) Function

The VCSE(t) is a time-variable function that describes the bias variations around the CCSE. It is a variable error component but different from RE. The RE changes unpredictably from measurement to measurement; meanwhile, VCSE(t) remains quasi-constant on a given day. The bias variations have unequal cycles, while the long-term mean of VCSE(t) is 0. Its values are not normally distributed.

VCSE(t) has 2 primary sources. Both are noninstrumental and specific. The variation in reagent quality follows a predictable pattern, and this variation is also predictable. After the calibration, we can predict the mean and bias variation from the old and new calibration parameters.

cbefore calibrationcafter calibrationFcalbefore calibrationFcalafter calibration

VCSE(t) is a mostly predictable phenomenon. We can correct it for a moment, but not definitively eliminate it. In repeatability conditions, the VCSE(t) is nonsense. The differences between CCSE (B_{RW}) (B_{RW}: long-term mean bias, measured in RW conditions, a constant), VCSE(t), and RE are presented in Table 7.

Table . Differences and similarities between the error components.

Criterion	RE[a]	VCSE(t)[b]	CCSE[c]
Predictability	Unpredictable	Yes, from the preceding data	Quasi-constant
Variability	Yes	Yes	No
Distribution caused	Normal	Non-Gaussian	Quasi-constant
Influence on the mean in reproducibility within laboratory conditions	Negligible (≈0)	Only after several complete cycles, it becomes negligible≈0	Yes
Calibration influence	Insignificant	It can be corrected, but not eliminated	Not significant
Corrections or correction factors, according to GUM	No effect	In the short term, yes, on long-term reappears	Yes
Measurable under repeatability conditions	s_r[d]	$B_r(t)$ includes VCSE(t)	No
Measurable under reproducibility within laboratory conditions	s_{RW}[e] includes s_r	s_{RW} includes s_{VCSE}[f]	B-RW

[a]RE: random error component.

[b]VCSE: variable component of the systematic error.

[c]CCSE: constant component of systematic error.

[d]s_r: SD measured in constant, repeatability conditions.

[e]s_{RW}: SD measured in variable, reproducibility within laboratory conditions.

[f]s_{VCSE}: SD calculable from the daily (run) mean, bias, or VCSE(t) values.

We cannot ignore the differences between VCSE(t), RE, and CCSE. If, and only if, we are conscious that both $B_r(t)$ and s_{RW} contain VCSE(t), it is not an erroneous practice to measure RE and VCSE(t) together and to include VCSE(t) in s_{RW}. The origins of the equations must be known, as well as the risk of redundant use.

Determination of the CCSE and the VCSE(t)

The determination of CCSE≡B-RW is possible using the control results and Equation 4. Such CCSE values only show the difference between the mean of control measurements and the target specified by the producer. We can obtain an absolute value of CCSE from the percent expressed EQA results. Due to the low number of measurements, the value has significant uncertainties.

The comparison between the 2 types of CCSE values is not the task of this study. A single mention: the difference between the 2 CCSE-s is predictably constant until we use the same control material. Another study is necessary to verify this prediction.

As a consequence of the constant difference, the VCSE(t) measured in internal QC and EQA predictably is the same; however, the statement needs confirmation. The accurate determination of the VCSE(t) function has a high cost-effectiveness ratio and negligible practical importance, mainly due to its short validity term. The computer-assisted estimation of the run means (Figure 5) is a promising solution, but needs a separate study to confirm its efficiency.

The same statement applies to the s_{VCSE}. To estimate s_{VCSE} using Equation 2 is more practical than calculating it from daily VCSE(t) values [27].

The increased s_{VCSE}/s_r ratio indicates wrong internal QC decisions (delayed calibrations); however, we can also use the $s_{RW}/s_{r\ ratio}$ without calculating s_{VCSE} [23].

The paramount importance of VCSE(t) and s_{VCSE} lies in the distinction between SD and bias types, not their absolute value. We do not need their accurate values; we do not make decisions based on them. These 2 parameters are always included in $B_r(t)$ or s_{RW}. However, we must be aware of where they are hidden. Highlighting VCSE(t) and s_{VCSE} in equations helps us avoid redundant use.

The Proposed Error Model and the Westgard Rules–Based Internal QC System

The original aim of this study was to draw attention to the neglected VCSE(t) and s_{VCSE}. The proposed new error model (Figure 6, Equations 2, 3, 5) also uncovers the weaknesses of the actual Westgard rules–based internal QC system. By distinguishing the biases measured in repeatability and reproducibility within conditions ($B_r[t]$ and B_{RW}), 2 sets of error parameters are obtained ($B_r[t]$ and s_r, respectively, for B_{RW} and s_{RW}). The link between them is VCSE(t) and s_{VCSE} (which are usually hidden in the $B_r(t)$ and s_{RW}). Avoiding redundant use by highlighting VCSE(t) and s_{VCSE} in equations is not the only advantage of the proposed error model. The non-Gaussian distribution of the VCSE(t) values explains the non-Gaussian distribution of the long-term QC data [3,4] and the significant monthly variability of s_{RW} [5], which contradicts the laws of the normal distribution. The Gauss-Laplace equation is valid only under constant repeatability conditions (if the mean remains constant). Therefore, s_r is the correct estimator of the Gaussian

σ parameter and the mean RE. While the sources of specific bias variability (reagent property and calibration parameter changes) are known [16,18,19,22,23], the sources of specific RE variability cannot be identified. All identifiable RE sources are linked to the inconsistent functionality of the instrument and, therefore, are constant (nonvariable) and nonspecific [38]. In contrast to s_{RW}, s_r is invariant within the limits of accuracy of the statistical methods (Vandra's unpublished data [38]).

The constant RE (s_r) questions the efforts of Westgard et al [1,8,39] to detect variations in RE. The primary objective of internal QC is to detect risky variations in bias, and, by definition, the bias between human interventions is predictable [14]. Anyway, according to Westgard JO [40], the QC rules cannot be applied across corrective actions. The objective change changes the way of thinking in QC. The focus is not on the immediate detection of unpredictable changes, but rather on following tendencies in bias to predict the moment when the run bias will reach a critical value.

There are 4 different mechanisms to reach a critical bias, imposing different decision strategies, because the QC rules (especially the cross-run rules: $R_{4\text{-}1S}$ and R_{10X}) have different efficiencies in each case.

1. Immediately after a calibration (Was the calibration successful?)
2. Constant bias in the case of a stable reagent (Is the new mean acceptable?)
3. Gradually increasing bias (in absolute values) in the case of an unstable reagent (When will the bias reach critical values?)
4. Unexpected shift in bias.

The immediate error detection is compulsory only in cases 1 and 4. In cases 2 and 3, bias is predictable. However, the QC system must be able to detect changes in the tendencies.

GRD Jones was the first to notice the difference between cases 1 and 4 [41], highlighting that in case 1, the cross-run rules ($R_{4\text{-}1S}$ and R_{10X}) cannot be applied due to a lack of data. However, he did not observe the hidden assumption in Westgard's calculations, which falsely assumes a constant bias in all runs. While focusing on immediate error detection in case 4, the calculations are based on case 2 (constant bias). If the cross-run rules detect a constantly critical bias, it indicates delayed, rather than immediate, error detection. In cases 3 and 4, the previous bias value is lower than in the last run, and the efficiency of the cross-run rules was overestimated.

In cases 2 and 3, the QC rules are applied repeatedly, increasing the efficiency of error detection. Instead of applying the $R_{1\text{-}3S}$ rule, the $R_{1\text{ of n-}3S}$ rule is used de facto. All runs are only accepted if neither of them violates the 3 SD decision limit.

The former observations impose the reevaluation of the efficiency of the Westgard rules in a subsequent study.

The Westgard rules are only correctly applied if the QC graphs are designed with σ or the correct estimator. As previously concluded, the correct estimator of the σ parameter and the mean RE is s_r, and Westgard's assumption that $s_{RW} \approx \sigma$ is false. Not else, but Westgard and Groth [39] acknowledged that:

> *The calculations based on computer simulations behind the power function graphs are made assuming within-run SD, while the graphs are designed with total SD.*

Considering the s_{RW}/s_r ratio, this results in an overestimation of the decision limits 1.5 - 2 times. Respecting Westgard's recommendations, intending to apply the $R_{1\text{-}3S}$ rule, de facto, we use the $R_{1\text{-}4.5S}$ or the $R_{1\text{-}6S}$ rule ($3s_{RW} \approx 4.5$ - $6s_r$.) This contradiction and overestimation explain the existence of the statistically impossible graphs observed in practice (mentioned in the Introduction).

Correcting the estimator of σ (from s_{RW} to s_r) requires recalculating all parameters that include SD in their equations: TE, MU, sigma metrics, the critical SE, not just a change in the design of the QC graphs. This means an entirely new QC system, using different rules and strategies.

Sounds bizarre, but according to calculations based on normal distribution tables, a correctly applied Westgard rules–based QC system (designing the graphs with σ) would be dysfunctional due to several false alarms. Despite the efforts to correct them, half of the monthly biases measured in the internal QC are around 1 s_{RW} or bigger, and two-thirds of them are bigger than $1s_r$. According to quintessential principle 2, it cannot be corrected by calibration for smaller biases than the average calibration error, questioning another assumption of Westgard et al [39]: the assumption of error-free calibrations. According to Vandra [38], the average calibration error is ≈1 - 2 s_r (consistent with the observed monthly biases). If such biases are incorrigible, the QC rules must avoid alarms in these cases. The correctly applied Westgard rules alarm in the first run only by exception if the bias is 0. The statement is not valid if $B>1s_r$ and the rules are applied in several runs.

Conclusions

This study is a theoretical one. It aims to draw the attention of the scientific community to the fact that the VCSE is a neglected phenomenon and a source of several errors. Because it is hidden in the inaccurately defined bias and the s_{RW}, there is the risk of its redundant use in equations. This study also aimed to uncover the primary sources of bias variations (both present in the literature in mosaic pieces), propose corrected equations, and describe the properties of the VCSE. Because several problems were uncovered, the proofs, based on computer simulations and real-life data for each issue, neither fit within the limits of a single study nor are consistent with the declared aims. To analyze them, subsequent studies will be necessary in the future. This study intends to be a starting point for building a new QC system based on a different error model, a different strategy, and a rule system. The theoretical foundations, description, proofs with computer simulation, and real-life data do not fit within the limits of this study.

The time variability of bias is a well-known but neglected phenomenon. A variable bias does not fit into the classical error model. If bias has variations, a question arises: Which bias is

being referred to? A new error model was obtained by (1) separating the bias into a constant and a variable subcomponent and (2) distinguishing between bias measured in repeatability and reproducibility within laboratory conditions. The error model is consistent with similar attempts found in the literature; however, it questions the theory of transformation of variable biases into random errors (based on an inaccurate definition of 'random' in VIM), which forces the VCSE into the Procrustes' bed of the old error model. The author proposed definitions consistent with the VIM 2.17 definition of the SE and abbreviations consistent with those used for SD (B-RW, $B_r(t)$).

The bias variability has 2 sources. Both are noninstrumental and specific to each measurement, and neither causes normally distributed biases. One is reagent instability, and the other is human intervention, including reagent changes and calibrations. Reagent instability causes gradually increasing, quasilinear biases, whereas calibrations result in alternation between constant periods with random shifts in the calibration parameters. Computer simulations and real-life QC data presented in this study support that these are real sources of bias variability.

The 2 phenomena occur simultaneously, resulting in sawtooth-like variations in bias. In the time frames between human interventions, the biases are predictable. However, they are hidden behind the noise of the RE. Without computer assistance, we can observe only significant shifts and drifts. For this reason, the increase of the SD in longer time frames was erroneously considered unpredictable, with an unknown cause (type b of unpredictable). An unpredictable bias contradicts its definition in VIM.

We must change our way of thinking in the QC by focusing on predictive actions instead of corrective ones.

The properties of the CCSE, the VCSE(t) function, and the RE differ, justifying the distinction between them. Accurately determining the SE subcomponents theoretically is possible; however, it has a high cost/effectiveness ratio. The significance of their separation is that they help us avoid the redundant use of the VCSE(t) classically hidden in $B_r(t)$ and s_{RW}.

Two sets of error parameters are obtained by separating biases measured in repeatability and reproducibility within laboratory conditions. We must determine the parameters under the same conditions we use them. UM calculations must be based on parameters determined under reproducibility within laboratory conditions, whereas internal QC decisions must be based on parameters determined under repeatability conditions. This conclusion is thought-provoking because it contradicts the recommendations for designing the Levey-Jennings graphs based on the SD calculated from long-term control data. In the meantime, the calculations behind the Westgard rules assume pure RE.

The actual Westgard rules–based internal QC system is not consistent with two quintessential principles valid in all sciences:

1. We must determine the parameters under the same conditions we use them.
2. A calibration cannot efficiently correct smaller biases than the mean calibration error.

The proposed error model uncovered several false assumptions behind the actual Westgard rules–based QC system.

1. The internal QC aims to detect variations in RE and SE. (RE is not variable.)
2. Bias variations are unpredictable. (Correct: between human interventions are predictable.)
3. The same rules are efficient in all cases. (Correct: there are 4 different situations of decisions, imposing different rules and strategies.)
4. Cross-run rules can be applied in immediate error detection. (Correct: they can be applied only with a delay.)
5. The estimator of the σ parameter and the measure of the mean RE is s_{RW} (Correct: it is s_r.)
6. QC graphs must be designed with $s_{RW.}$ (Correct: with s_r, highlighting the incorrigible biases.)
7. Calibrations are error-free, and all biases are correctable by calibration. (Correct: smaller biases than $1 - 2s_r$ are incorrigible.)

The false assumptions 6 and 7 cause 2 compensating errors. The compensation explains the long-term success of the Westgard rules. If we use s_{RW} in the design of the Levey-Jennings graphs, we use larger, increased decision limits, de facto applying different rules (eg, the $R_{1\text{-}5S}$ rule instead of the intended $R_{1\text{-}3S}$). As a consequence, the alarms for incorrigible biases become less frequent. However, this compensation is not accurate. The observed statistically impossible QC graphs sustain the overestimation of the RE by the s_{RW}.

Based on the proposed error model, correcting the former false assumptions, and considering the 4 different decision situations, the Westgard rules–based QC system must be mathematically reevaluated. It can be predicted that patching it is not a solution, and a new QC system is necessary, based on the $s_{r,}$ and the avoidance of alarms in the case of incorrigible biases.

The proposed error model also suggests corrections to the MU equations. MU is a long-term parameter, and therefore, its equation must be based on long-term parameters. The uncertainty of the inaccurately defined bias (Which one?) must be substituted with the uncertainty of the long-term mean bias, measured in reproducibility within laboratory conditions (U(B-RW)), and must be considered the uncertainty caused by the variability of s_{RW}, substituting it with its maximal value in the MU equation.

Furthermore, the proposed error model, together with quintessential principle 1 (that all parameters must be determined under the same conditions under which they are used), explains why the more correct MU theory cannot substitute for TE in internal QC decisions. MU is a long-term parameter, while internal QC decisions are made under repeatability conditions.

Acknowledgments

The author thanks Dr Prof Marius Măru teri for the initial reading, valuable advice, and constructive critiques that helped improve the study, as well as the reviewers' critical opinions. To this study, no other persons contributed except the author. The author created all images and tables. The author attests that this manuscript did not use generative artificial intelligence (AI) technology to generate figures, ideas, data, or other informational content. AI was used only for grammar correction and for unintentional plagiarism detection. To assist with the language correction, the author used the following Grammarly AI prompts: “Improve it” and “Find synonyms.”

Data Availability

All computer simulation files were uploaded as Multimedia Appendices 1 and 2 (in Excel format). The data, which constituted the basis of the real-life data graphs, were also uploaded as Multimedia Appendices 3 and 4 (Excel files). The latter data source is the quality control results obtained in the Brasov County Clinical Hospital for Urgencies (Romanian abbreviation: SCJUBv), part of a protected database; therefore, these cannot be made available. The author did not use patient data in this study. In the real-life examples, reference materials produced by Roche were used.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Random values generation and shift and drift computer simulation.
[XLSX File, 925 KB - xmed_v7i1e49657_app1.xlsx]

Multimedia Appendix 2
Differences in calibrations and shift and drift computer simulations (protected file).
[XLSX File, 407 KB - xmed_v7i1e49657_app2.xlsx]

Multimedia Appendix 3
Real-life data for glucose. The influence of reagent degradation on the bias variation.
[XLSX File, 277 KB - xmed_v7i1e49657_app3.xlsx]

Multimedia Appendix 4
Real-life data for magnesium. Shift caused by calibration.
[XLSX File, 48 KB - xmed_v7i1e49657_app4.xlsx]

References

1. Westgard JO, Barry PL, Hunt MR, Groth T. A multi-rule Shewhart chart for quality control in clinical chemistry. Clin Chem 1981 Mar;27(3):493-501. [doi: 10.1093/clinchem/27.3.493] [Medline: 7471403]
2. Stahl S. The evolution of the normal distribution. Math Mag 2006 Apr;79(2):96-113. [doi: 10.1080/0025570X.2006.11953386]
3. Badrick T. Biological variation: understanding why it is so important? Pract Lab Med 2021 Jan;23:e00199. [doi: 10.1016/j.plabm.2020.e00199] [Medline: 33490349]
4. Katayev A, Fleming JK. Past, present, and future of laboratory quality control: patient- based real-time quality control or when getting more quality at less cost is not wishful thinking. J Lab Precis Med 2020;5:28-28. [doi: 10.21037/jlpm-2019-qc-03]
5. Kumar BV, Mohan T. Sigma metrics as a tool for evaluating the performance of internal quality control in a clinical chemistry laboratory. J Lab Physicians 2018;10(2):194-199. [doi: 10.4103/JLP.JLP_102_17] [Medline: 29692587]
6. [VIM3] 2.19 random measurement error. Joint Committee for Guides in Metrology. 2025. URL: https://jcgm.bipm.org/vim/en/2.19.html [accessed 2025-06-30]
7. Krystek M. Calculating Measurement Uncertainties: Beuth Verlag GmbH; 2016.
8. Westgard JO, Carey RN, Wold S. Criteria for judging precision and accuracy in method development and evaluation. Clin Chem 1974 Jul;20(7):825-833. [doi: 10.1093/clinchem/20.7.825] [Medline: 4835236]
9. How to calculate your long term bias for your uncertainty calculation? Weqas. 2020 Jul 13. URL: https://www.weqas.com/download/how-to-calculate-your-long-term-bias-for-your-uncertainty-calculation/ [accessed 2022-05-29]
10. Tholen DW. Evaluation of the linearity of quantitative measurement procedures: a statistical approach; approved guideline. NCCLS. 2003 URL: https://mdcpp.com/doc/standard/NCCLSEP6-A-2003.pdf [accessed 2026-01-28]
11. JCGM GUM-6:2020. Bureau International des Poids et Mesures (BIPM). URL: https://www.bipm.org/documents/20126/2071204/JCGM_GUM_6_2020.pdf/d4e77d99-3870-0908-ff37-c1b6a230a337 [accessed 2025-05-06]

12. JCGM GUM 100:2008 evaluation of measurement data — guide to the expression of uncertainty in measurement. Bureau International des Poids et Mesures (BIPM). 2008 Jan 1. URL: http://www.bipm.org/en/publications/guides/gum.html [accessed 2025-06-29]
13. Leito I. Validation of liquid chromatography mass spectrometry (LC-MS) methods. University of Tartu. URL: https://sisu.ut.ee/lcms_method_validation/51-Bias-and-its-constituents [accessed 2024-05-29]
14. [VIM3] 2.17 systematic measurement error. Joint Committee for Guides in Metrology. 2025. URL: https://jcgm.bipm.org/vim/en/2.17.html [accessed 2025-06-30]
15. Shewhart WA. Economic quality control of manufactured product. Bell Syst Tech J 1930;9(2):364-389. [doi: 10.1002/j.1538-7305.1930.tb00373.x]
16. Marquis P. Common misconceptions in medical laboratory quality control. : Service de Biochimie, Centre hospitalier régional Metz – France URL: https://files.secure.website/wscfus/4091441/31621936/misconceptions.pdf [accessed 2026-01-28]
17. Eisenhart C. Realistic evaluation of the precision and accuracy of instrument calibration systems. J Res Natl Bur Stan Sect C Eng Instr 1963 Apr;67C(2):161. [doi: 10.6028/jres.067C.015]
18. Haeckel R, Schneider B. Detection of drift effects before calculating the standard deviation as a measure of analytical imprecision. Clin Chem Lab Med 1983;21(8):491-498. [doi: 10.1515/cclm.1983.21.8.491]
19. Krouwer JS. Setting performance goals and evaluating total analytical error for diagnostic assays. Clin Chem 2002 Jun;48(6 Pt 1):919-927. [doi: 10.1093/clinchem/48.6.919] [Medline: 12029009]
20. Kadis R. Evaluation of measurement uncertainty in analytical chemistry: related concepts and some points of misinterpretation. ResearchGate. 2008. URL: https://www.researchgate.net/publication/277054223_Evaluation_of_measurement_uncertainty_in_analytical_chemistry_related_concepts_and_some_points_of_misinterpretation [accessed 2025-03-07]
21. Theodorsson E, Magnusson B, Leito I. Bias in clinical chemistry. Bioanalysis 2014;6(21):2855-2875. [doi: 10.4155/bio.14.249] [Medline: 25486232]
22. Magnusson B, Näykki T, Hovind H, Krysell M, Sahlin E. Handbook for calculation of measurement uncertainty in environmental laboratories (NT TR 537 - edition 4). NORDTEST. 2017 Nov 29. URL: http://www.nordtest.info/wp/2017/11/29/handbook-for-calculation-of-measurement-uncertainty-in-environmental-laboratories-nt-tr-537-edition-4/ [accessed 2025-03-07]
23. Mackay M, Hegedus G, Badrick T. Assay stability, the missing component of the error budget. Clin Biochem 2017 Dec;50(18):1136-1144. [doi: 10.1016/j.clinbiochem.2017.07.004] [Medline: 28733188]
24. Oosterhuis WP, Bayat H, Armbruster D, et al. The use of error and uncertainty methods in the medical laboratory. Clin Chem Lab Med 2018 Jan 26;56(2):209-219. [doi: 10.1515/cclm-2017-0341] [Medline: 28796637]
25. Vandra AB. Incertitudini… în lumea incertitudinii. Deplasarea [Uncertainties... in the world of uncertainties. The bias]. Revista română de laborator medical 2014 Sep.
26. Gauss CF. Bestimmung der Genauigkeit der Beobachtungen [determining the accuracy of the observations]. Z Astron Verw Wiss [J Astron Relat Sci] 1816;1:187-197.
27. Vandra AB. Reevaluation of the variable component of the systematic error calls for paradigm change in clinical laboratory quality control. Health Systems and Quality Improvement. Preprint posted online on May 28, 2023. [doi: 10.1101/2023.05.24.23290382]
28. Magnusson B, Ellison SLR. Treatment of uncorrected measurement bias in uncertainty estimation for chemical measurements. Anal Bioanal Chem 2008 Jan;390(1):201-213. [doi: 10.1007/s00216-007-1693-1] [Medline: 18026721]
29. White GH, Farrance I, AACB Uncertainty of Measurement Working Group. Uncertainty of measurement in quantitative medical testing: a laboratory implementation guide. Clin Biochem Rev 2004;25(4):S1-24. [Medline: 18650962]
30. Pang R. A guide on how to implement internal quality control (IQC) HKAML 2024. ResearchGate. 2024 Mar 27. URL: https://www.researchgate.net/publication/379333346 [accessed 2025-07-01]
31. Westgard S. Uncertainty in how to calculate measurement uncertainty: different approaches for incorporating effects of clinically significant bias. WestgardQC. 2023. URL: https://westgard.com/essays/iso/uncertainty-in-uncertainty.html [accessed 2025-05-05]
32. White GH. Basics of estimating measurement uncertainty. Clin Biochem Rev 2008 Aug;29 Suppl 1(Suppl 1):S53-S60. [Medline: 18852859]
33. Kristensen GBB, Meijer P. Interpretation of EQA results and EQA-based trouble shooting. Biochem Med (Zagreb) 2017 Feb 15;27(1):49-62. [doi: 10.11613/BM.2017.007] [Medline: 28392726]
34. Vlašić Tanasković J, Coucke W, Leniček Krleža J, Vuković Rodriguez J. Peer groups splitting in Croatian EQA scheme: a trade-off between homogeneity and sample size number. Clin Chem Lab Med 2017 Mar 1;55(4):539-545. [doi: 10.1515/cclm-2016-0284] [Medline: 27658147]
35. Toacşe G, Toacşe AM. Controlul de Calitate Şi Validarea Metodelor Analitice Cantitative: Pentru Laboratoarelor Medicale Bucureşti [Quality Control and Validation of Quantitative Analytical Methods: For the Use of Clinical Laboratories Bucharest]: Editura Tehnică; 2010.
36. Badrick T. The quality control system. Clin Biochem Rev 2008 Aug;29 Suppl 1(Suppl 1):S67-S70. [Medline: 18852861]

37. Burtis CA, Ashwood ER, editors. Tietz Textbook of Clinical Chemistry 2: W.B. Saunders; 1994.
38. Vandra AB. Calibration error, a neglected error source in the clinical laboratory quality control. EJIFCC 2025 Dec;36(4):443-451. [Medline: 41459182]
39. Westgard JO, Groth T. Power functions for statistical control rules. Clin Chem 1979 Jun;25(6):863-869. [doi: 10.1093/clinchem/25.6.863] [Medline: 445821]
40. QC - the multirule interpretation. WestgardQC. URL: https://www.westgard.com/lessons/basic-qc-practices-l/31-lesson18.html [accessed 2025-07-01]
41. Jones GRD. Reevaluation of the power of error detection of Westgard multirules. Clin Chem 2004 Apr;50(4):762-764. [doi: 10.1373/clinchem.2003.025585] [Medline: 15044336]

Abbreviations

B_{RW}: long-term mean bias, measured in RW conditions, a constant
CCSE: constant component of systematic error
CV: coefficient of variation, the SD expressed as a percent of the mean of measurements
CV_r: CV measured in constant, repeatability conditions
CV_{RW}: CV measured in variable, reproducibility within laboratory conditions
EQA: external quality assessment
IQC: internal quality control
QC: quality control
RE: random error component
SE: systematic error component
s_r: SD measured in constant, repeatability conditions
s_{RW}: SD measured in variable, reproducibility within laboratory conditions
s_{VCSE}: the SD calculable from the daily (run) mean, bias, or VCSE(t) values
TE: total measurement error
UM: uncertainty of measurement
VCSE: variable component of the systematic error
VIM: International Vocabulary of Metrology

Edited by T Leung; submitted 05.Jun.2023; peer-reviewed by E Theodorsson, Anonymous; revised version received 09.Jul.2025; accepted 30.Nov.2025; published 27.Feb.2026.

Please cite as:
Vandra AB
Investigating the Variable Component of the Systematic Error, a Neglected Error Parameter: Theoretical Reevaluation Study
JMIRx Med 2026;7:e49657
URL: https://xmed.jmir.org/2026/1/e49657
doi:10.2196/49657

Awareness, Experiences, and Attitudes Toward Preprints Among Medical Academics: Convergent Mixed Methods Study

Mustafa Sevim[1*], Dr med; Burak Karamese[2*]; Zafer Alparslan[2*]

[1]Department of Physiology, School of Medicine, Marmara University, Başıbüyük Yolu No: 9 D:2, Istanbul, Turkey

[2]School of Medicine, Marmara University, İstanbul, Turkey

[*]all authors contributed equally

Corresponding Author:
Mustafa Sevim, Dr med
Department of Physiology, School of Medicine, Marmara University, Başıbüyük Yolu No: 9 D:2, Istanbul, Turkey

Related Articles:

Companion article: https://www.medrxiv.org/content/10.1101/2025.04.09.25325524v1

Companion article: https://med.jmirx.org/2026/1/e95736

Companion article: https://med.jmirx.org/2026/1/e95737

Companion article: https://med.jmirx.org/2026/1/e95735

Abstract

Background: Preprints—scientific manuscripts shared publicly prior to formal peer review—are gaining momentum across academic disciplines. However, their adoption in clinical and biomedical sciences remains limited, particularly in countries where traditional publishing norms prevail. Editorial ambiguity and a lack of national policy further complicate their use.

Objective: This study aimed to assess the awareness, experiences, and attitudes of medical academics at Marmara University School of Medicine toward preprints and to explore the editorial landscape through both journal editor feedback and a review of journal-level preprint policies.

Methods: A cross-sectional survey was conducted with 103 medical faculty members. The questionnaire included demographic questions, Likert scale items, and multiple-choice items assessing knowledge, familiarity, and attitudes toward preprints, as well as open-ended items to explore concerns. A "preprint test score" (0‐4) was developed to quantify objective knowledge. Subgroup analyses were conducted by age (<40 vs ≥40 y) and academic discipline (basic vs clinical sciences). Additionally, all responses to open-ended questions from journal editors and 118 biomedical journals were manually reviewed for their stated stance on preprints and article processing charges (APCs). A convergent mixed methods design was used, combining a structured survey, thematic analysis of open-ended responses and editorial feedback, and a document-based review of biomedical journal policies.

Results: Only 42.9% (n=34) of participants reported familiarity with the concept of preprints, and 13% (n=10) had previously published on a preprint server. Misconceptions about ethics, peer review, and compatibility with journal policies were common. Subgroup analysis revealed that older participants scored higher on the "preprint test" (mean 2.20, SD 1.31 vs mean 1.97, SD 1.60) and had more experience with preprint publishing (1/40, 2.5% of younger participants; 7/29, 24.1% of older participants). Further, younger academics expressed less openness toward future use (n=7, 17.5% in the younger group; n=8, 27.6% in the older group). Clinical faculty were generally more hesitant than basic science faculty, although both groups raised concerns about the academic recognition of preprints. Editorial responses reflected a mix of cautious endorsement and skepticism. Among the 118 biomedical journals reviewed, most lacked clear preprint policies, while a small number either explicitly prohibited or permitted them.

Conclusions: There is limited awareness and cautious engagement with preprints among medical academics and editors in Türkiye. Generational and discipline-based differences further influence knowledge and attitudes. The lack of clear editorial guidance from biomedical journals may reinforce academic uncertainty. Tailored educational initiatives, transparent journal policies, and institutional support will be essential to foster a more open and inclusive scientific publishing environment.

(JMIRx Med 2026;7:e78139) doi:10.2196/78139

KEYWORDS

preprint; medical academics; publishing attitudes; editorial policies; survey

Introduction

Preprints—manuscripts publicly shared prior to peer review—have transformed the pace and openness of scholarly communication. Widely adopted in fields such as physics and computer science for decades, preprints enable rapid dissemination, open peer commentary, and broader visibility of research findings [1]. In recent years, the biomedical community has increasingly engaged with preprint platforms, particularly during public health crises such as the COVID-19 pandemic. For instance, between June 2020 and June 2022, the US National Library of Medicine made more than 3300 National Institutes of Health–funded COVID-19 preprints accessible in PubMed Central, marking a pivotal shift toward preprint integration in mainstream biomedical publishing [2].

Despite this global momentum, the adoption of preprints in clinical and medical sciences remains uneven [3]. In countries like Türkiye, where academic evaluation systems and journal structures still emphasize traditional peer-reviewed publication, the concept and utility of preprints are often misunderstood or undervalued. Anecdotal observations suggest hesitancy among medical faculty, fuelled by concerns about plagiarism, duplication, and lack of recognition in academic promotion criteria.

Journal editors also play a crucial role in shaping scholarly norms. Editorial policies on preprints vary widely across journals: while some encourage their use, others either prohibit them or do not explicitly mention them at all [4-7]. The absence of a clear preprint policy creates uncertainty for authors and may contribute to low adoption, particularly among early-career researchers concerned about publication eligibility [8].

While international studies have explored general attitudes toward preprints [3,9-11], little is known about how these perspectives vary within academic subgroups. Factors such as career stage and departmental discipline may influence both knowledge and perception. For instance, basic science researchers are often more open to experimentation with publishing models, whereas clinical academics may prioritize peer-reviewed evidence with clear implications for practice [12]. Similarly, younger faculty may view preprints as tools for early visibility and career advancement, while more senior academics may adhere to traditional notions of scholarly validation and prestige.

To date, no systematic assessment has examined the knowledge, attitudes, and editorial perspectives regarding preprints within Türkiye's medical academic community. To address this gap, this study used a mixed methods design, integrating three data sources: a cross-sectional survey of medical academics at Marmara University School of Medicine; a document-based review of preprint and article processing charge (APC) policies from Turkish biomedical journals; and a descriptive analysis of qualitative feedback from journal editors.

In addition to characterizing general patterns, we examine subgroup differences by age and academic discipline to uncover nuanced barriers and opportunities for preprint adoption in the evolving landscape of scientific communication.

Methods

Overview

This study involved a cross-sectional survey design conducted at Marmara University School of Medicine in İstanbul, open-ended questions directed at journal editors, and document-based content analysis (convergent mixed methods). The goal was to evaluate the awareness, knowledge, and attitudes of medical academics toward preprints and to explore perceived barriers to preprint use.

Overall, this study harnessed three different data sources to evaluate this preprint issue:

- An online survey for medical academics at the Marmara University School of Medicine that included both quantitative and qualitative (open-ended questions) elements.
- A comprehensive review of editorial perspectives and policies of Turkish biomedical journals regarding preprints.
- Open-ended answers obtained from the editors of those biomedical journals regarding preprint policies and attitudes.

Participant Recruitment and Data Collection

A structured online survey (closed survey) was created using SurveyMonkey to control all technical issues and distributed via institutional email lists and internal professional networks between April and July 2024. At the beginning of the survey, participants were informed about the survey and informed consent was obtained. The survey targeted medical academics from a variety of departments and academic ranks at Marmara University School of Medicine. We reached out to all 1529 medical academics. No personal identifiers were collected, and all responses were anonymized and aggregated for analysis.

Although the total number of participants was 108, not all respondents answered every question. Some skipped certain items, particularly in the later sections of the survey. As a result, the number of responses varies across different variables, and this is reflected in the sample sizes reported in the Results section.

Survey Instrument

The survey included demographic items (eg, age, academic title, department), multiple-choice questions, and Likert scale questions assessing familiarity with preprints, previous use, attitudes toward preprints and peer review, and expectations for scientific quality and open-ended questions capturing perceived barriers and concerns related to preprint use.

To quantify objective knowledge of preprints, a "preprint test score" was generated from 4 multiple-choice questions. Participants received one point for each correct response,

resulting in a total score ranging from 0 to 4. A higher score indicated higher knowledge of preprints. The responses to the relevant part of the survey (Question 9: "Tick the option you think is correct") were used to calculate the preprint test score. The whole survey form can be found in Multimedia Appendix 1.

Subgroup Analyses

To explore differences in preprint engagement and perceptions, participants were stratified into subgroups based on the following:

- Age: Participants were divided into two groups based on age—those younger than 40 years and those 40 or older. This age threshold was chosen for two primary reasons. First, it closely reflects the central tendency of our sample's age distribution (mean 39.56, median 36, range 23‐73 years). Second, within the Turkish academic context, 40 represents a significant career milestone, often coinciding with the transition to an assistant professorship.
- Academic discipline: Basic sciences (eg, physiology, microbiology) versus clinical sciences (eg, internal medicine, surgery).

Subgroup comparisons were made for knowledge scores, attitudes, and future intent to use preprints, allowing the identification of generational and discipline-based trends.

Editorial Perspectives From Turkish Biomedical Journals

To gather complementary insights into institutional attitudes toward preprints, we reviewed all journals indexed in the Web of Science (InCites dataset and Emerging Sources Citation Index), filtered for Türkiye as the country of publication and covering the time period from 2019 to 2023. From an initial list of 280 journals, 264 remained after excluding duplicates, inaccessible websites, and journals with unclear policies. The 2-year impact factors and Journal Citation Index (JCI) quartiles were obtained as well.

These journals were manually categorized into biomedical and nonbiomedical fields based on their scope and published content according to Web of Science categories. Journals within the disciplines of medical sciences, pharmacology, biology, veterinary sciences, and nursing were classified as biomedical. Based on this classification, we identified 118 biomedical journals indexed in the dataset and based in Türkiye (as of April 2025) (Multimedia Appendix 2).

Editors of these journals were contacted via email and invited to respond to three open-ended questions (Multimedia Appendix 3): (1) their journal's current stance on preprints, (2) the rationale behind that stance, and (3) their views on the future role of preprints in academic publishing.

The email was sent to the editors-in-chief of all biomedical journals, and a total of 7 editors responded. Given the limited number of responses, the data were summarized descriptively rather than subjected to formal thematic analysis.

Journal Policy Review

In parallel, the same 118 biomedical journals were reviewed to assess their formal policies on preprints and APCs. Policy information was manually collected by examining publicly available sections of the journals' websites, including "Instructions for Authors," "Editorial Policy," and "Ethical Guidelines."

This assessment was conducted in two rounds—first in February 2024 and again in April 2025—to track any changes in policy over time.

Journals were categorized as follows:

- Preprint policy: allowed, prohibited, or not mentioned.
- APC policy: obligatory, free, or case-by-case (where charges depend on article type or other conditions).

This analysis provided insight into the editorial infrastructure surrounding preprints in Türkiye and helped contextualize how journal-level policies may influence researchers' behavior and perceptions.

Study Design and Data Analyses

This study used a convergent mixed methods design integrating quantitative survey data, qualitative insights, and document-based content analysis to explore medical academics' awareness and attitudes regarding preprints.

A cross-sectional online questionnaire included demographic questions, Likert scale questions, and multiple-choice items. Descriptive statistics (frequencies, percentages, means, medians, SDs) were used, and no inferential statistical testing was conducted. This survey was reported in accordance with the Checklist for Reporting Results of Internet E-Surveys (CHERRIES) [13] (Checklist 1).

Quantitative comparisons of journal policies were made using frequency counts and visualized via bar plots and heatmaps, with reference to impact metrics (eg, JCI quartiles).

Open-ended responses within the survey were analyzed using pattern-based thematic analysis. Commonly expressed concerns were coded inductively to identify recurrent barriers and perceptions regarding preprint use. Responses were grouped into themes such as plagiarism concerns, lack of academic recognition, policy confusion, and ethical ambiguity.

Editorial perspectives were obtained through open-ended email queries sent to biomedical journal editors. These responses were descriptively summarized to illustrate common institutional views and infrastructure limitations regarding preprint adoption.

Findings from the three data sources were integrated during interpretation to identify convergence and divergence. Quantitative trends were contextualized with qualitative themes and policy landscape shifts, enabling a holistic understanding of both individual attitudes and institutional structures shaping preprint practices in Türkiye.

Ethical Considerations

The study was approved by the Marmara University Faculty of Medicine Non-Drug and Medical Device Research Ethics

Committee (protocol code 09.2024.600; May 17, 2024). Informed consent was obtained from all participants prior to enrollment. To ensure participant confidentiality, all datasets were thoroughly de-identified for preceding statistical evaluation, and the study was conducted in accordance with the principles of the Declaration of Helsinki. No compensation was provided to participants.

Results

Journal Policy Review

As part of our review of Turkish biomedical journal policies, we analyzed the trends in preprint and APC policies at two time points: February 2024 and April 2025. This dual time point approach aimed to assess how Turkish biomedical journals are evolving in response to global shifts in open science practices and publication economics.

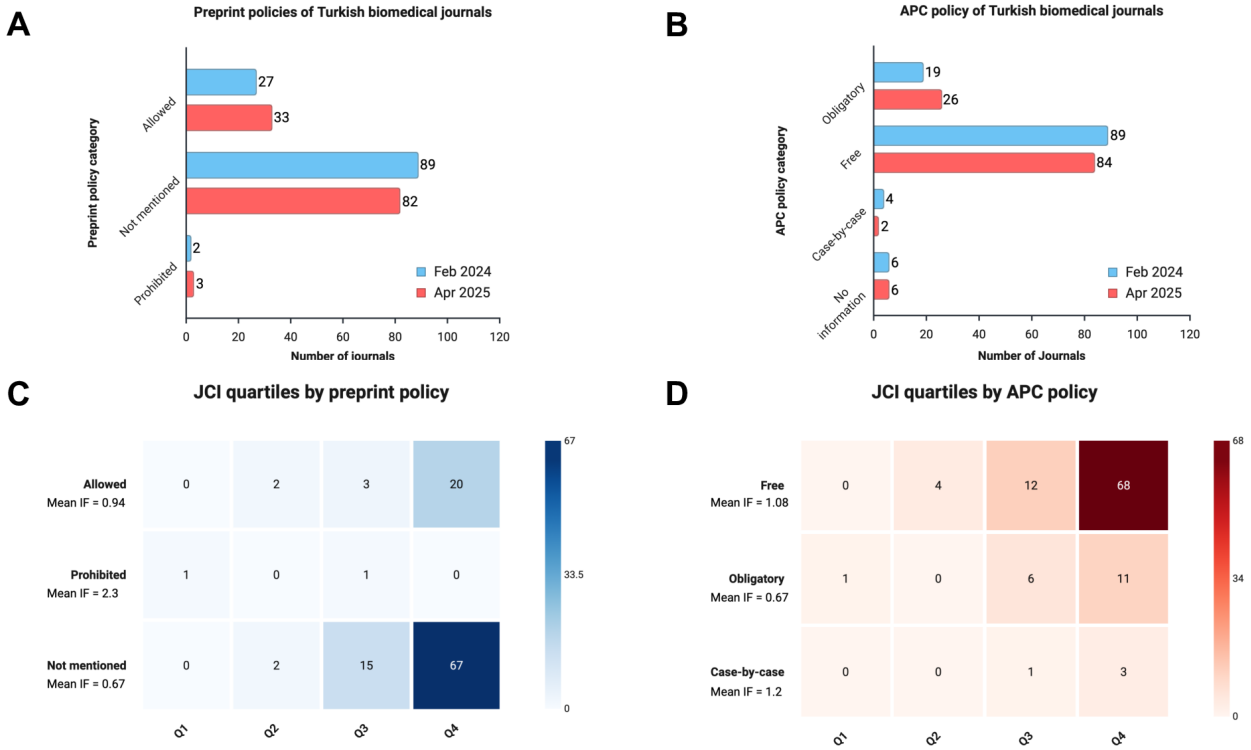
Preprint Policies

Preprint policies were categorized into three groups:

- Allowed: Journals explicitly welcome or permit submissions that were previously posted as preprints.
- Prohibited: Journals clearly disallow submissions that have appeared as preprints.
- Not mentioned: No reference to preprints could be found on the journal's official website.

Although the majority of journals still do not provide explicit guidance on preprints, our comparison over time revealed a modest but meaningful shift toward acceptance. Between February 2024 and April 2025, the number of journals explicitly allowing preprints increased from 27 to 33, while those with no policy decreased from 89 to 82. One additional journal began explicitly prohibiting preprints, increasing that category from 2 to 3 journals. These findings suggest a gradual trend toward policy transparency and a slow but positive normalization of preprint culture among Turkish biomedical journals (Figure 1A).

Figure 1. Preprint policy and article processing charge (APC) breakdown of Turkish biomedical journals. (A) Change in declared preprint policies of Turkish biomedical journals between February 2024 and April 2025. Journals that explicitly allowed preprints increased from 27 to 33, while those with no policy slightly decreased. (B) Change in APC policies of the same journals over the same period. A small rise in journals requiring obligatory APCs was observed. (C) Heatmap showing the distribution of journals across Journal Citation Index (JCI) quartiles by preprint policy, alongside mean journal impact factor (IF). Journals in all categories were predominantly concentrated in lower quartiles, particularly Q4; the prohibited group had the highest mean IF, but this category included only a very small number of journals. (D) Heatmap showing JCI quartile distribution by APC policy. Journals with case-by-case APCs had the highest mean IF, followed by free journals, whereas obligatory APC journals had a lower mean IF; all APC policy groups were concentrated mainly in Q3 and Q4.



The implications of these policy differences are further reflected in journal performance metrics. As shown in the heatmap (Figure 1C), journals that allow preprints had a mean impact factor of 0.94, compared with 0.67 for journals that did not mention a preprint policy, whereas the prohibited category had the highest mean impact factor (2.3), although this was based on a very small sample.

APC Policies

We also examined APC policies, grouping them into the following categories:

- Obligatory: Journals that always charge a fee for publication.
- Free: No publication charges to authors.

- Case-by-case: Charges apply only under certain conditions (eg, article type, page length).
- No information: No public declaration of APC policy found.

Between the two assessment periods, we noted a slight increase in journals with obligatory APCs (from 19 to 26), accompanied by a minor decrease in "free" journals (from 89 to 84). This indicates that APCs are becoming more common among Turkish biomedical journals, potentially impacting submission decisions, especially for early-career or unfunded researchers (Figure 1B).

When examining journal performance relative to APC policy, journals with case-by-case APC models showed the highest mean impact factor (1.2), followed by free journals (mean IF=1.08), whereas journals with obligatory APCs had a lower mean IF (mean IF=0.67) (Figure 1D). This suggests that journals with structured APC policies may be more established or competitive in the academic publishing ecosystem.

Results of the Survey

Overview

A total of 103 medical academics participated in the study. Among those who reported sex (n=98), 51% were female and 49% were male. Age distribution (mean 39.56, SD 12.64, median 36 years) ranged from early 20s to over 70, with a wide representation from junior researchers to senior professors. In terms of academic titles, professors (28.1%) and residents (29.2%) made up the largest groups (Table 1).

Table . Demography of the participants (N=103).

Characteristic and categories		Values
Gender, n (%)		
	Female	50 (51)
	Male	48 (49)
Age (years), mean (SD)		39.56 (12.64)
Academic title, n (%)		
	Professor	27 (28.1)
	Associate professor	8 (8.3)
	Assistant professor	15 (15.6)
	Lecturer	7 (7.3)
	Resident (medical specialty)	28 (29.2)
	Other	11 (11.5)
Department, n (%)		
	Basic medical sciences	29 (30.2)
	Internal medical sciences	54 (56.3)
	Surgical medical sciences	13 (13.5)
	General practitioner	0 (0)
Scientific publications (last 5 y), n (%)		
	0	15 (15.6)
	1 - 5	37 (38.5)
	6 - 10	18 (18.8)
	11 - 15	7 (7.3)
	16 - 20	5 (5.2)
	≥20	14 (14.6)

Awareness of Preprints

A substantial portion of participants lacked awareness of preprints. Only 43% (n=36/84) of respondents reported familiarity with the term (mean test score of 2.76), while 21.4% (n=18/84; mean test score of 1.0) had never heard of it. Finally, 36% (n=30/84) had heard of the term but were not familiar with preprints (mean test score of 1.85).

Knowledge of Preprints

To quantify participants' objective knowledge of preprints, a "preprint test score" was calculated based on responses to 4 multiple-choice questions. Participants received 1 point for each correct answer, resulting in a total score ranging from 0 to 4.

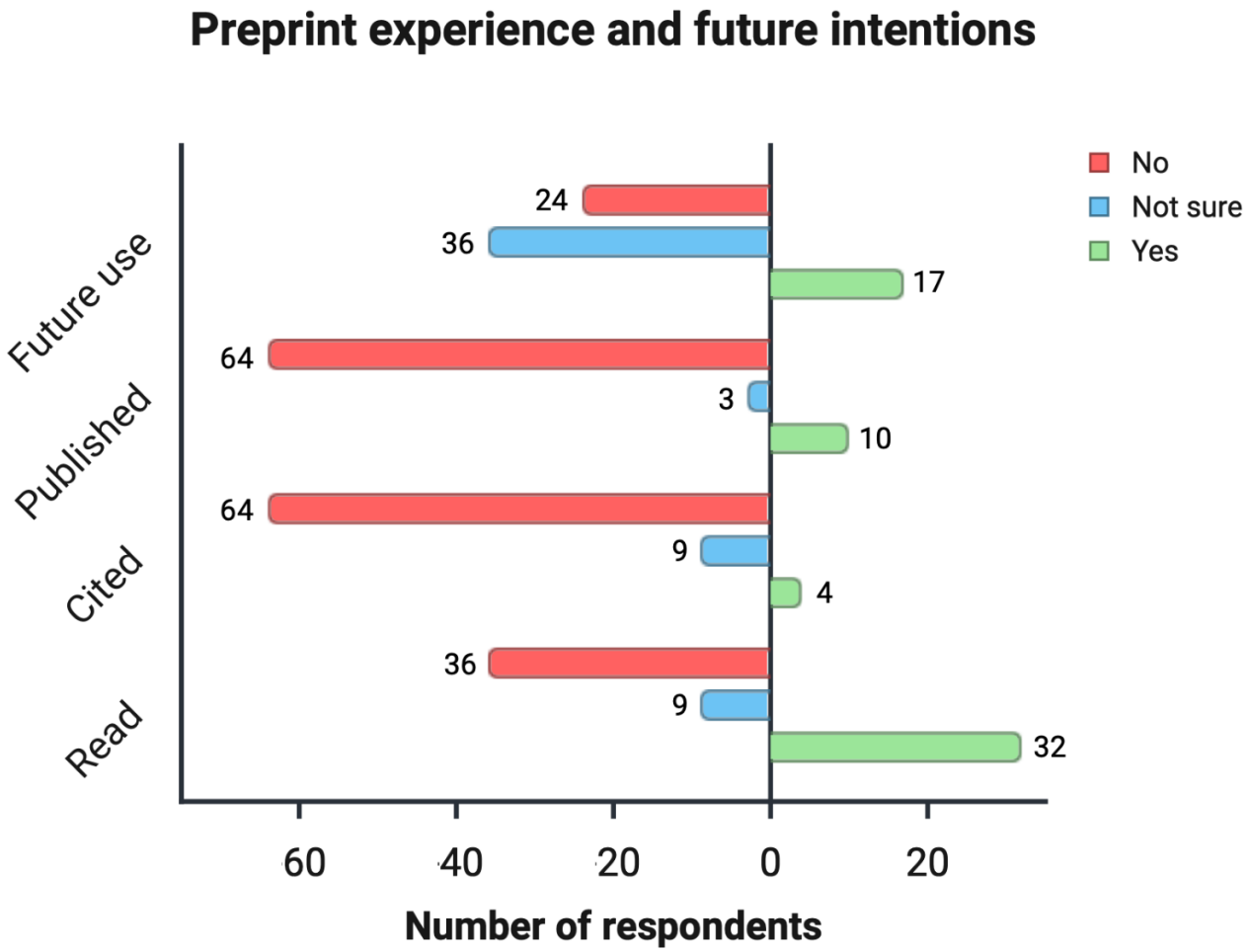
Overall, the mean preprint test score across all respondents was moderate (mean 2.07), with variation seen across both age groups and academic disciplines. Participants who reported being familiar with the term "preprint" scored highest (mean

2.76), while those who had never heard of the term scored the lowest (mean 1.00), suggesting internal validity of the test score.

Experience With Preprints

Engagement with preprints was low across all indicators. Only 10 participants had published on a preprint server, 32 had read a preprint, and just 4 had cited a preprint in their scientific writing (Figure 2).

Figure 2. Experience with preprints among medical academics. Horizontal bar chart showing the distribution of responses regarding participants' preprint-related behaviors and intentions, including whether they have read, cited, published, or plan future use of preprints. Responses are grouped by "yes," "no," and "not sure," highlighting overall low levels of direct engagement with preprints.



Attitudes Toward Future Use

When asked whether they would consider publishing a future manuscript as a preprint, only 22.1% (17/77) said yes, while 46.7% (n=36/77) were unsure and 31.2% (n=24/77) said no (Figure 2).

Subgroup Analyses

Subgroup Analysis by Age

Participants were categorized into two age groups: younger (<40 y, n=40) and older (≥40 y, n=29). The mean ages were 30.85 and 51.86 years, respectively. The older group demonstrated greater familiarity with preprints and higher preprint test scores (mean 2.20, SD 1.31 vs mean 1.97, SD 1.60). While only 2.5% (n=1) of younger participants had published a preprint, 24.1% (n=7) of older participants had done so. Further, younger participants showed less openness to future use, with 17.5% (n=7) considering preprint publication compared to 27.6% (n=8) in the older group, and a larger proportion remaining undecided (n=18, 62.5%).

Interestingly, younger participants expressed more favorable attitudes toward the role of preprints in scientific development and reforming the peer review system, whereas older participants valued conventional indicators of study quality such as citation counts and journal prestige. These differences are illustrated in Table 2.

Table . Attitudes and perceptions toward preprints by age group (N=69).[a]

Questions	<40 y (n=40)			≥40 y (n=29)		
	Agree	Neither agree nor disagree	Disagree	Agree	Neither agree nor disagree	Disagree
Preprints contribute to scientific development	29 (72.5)	11 (27.5)	0 (0)	13 (44.8)	12 (41.4)	4 (13.8)
The way preprints change the peer-review and editorial process is favorable for the future of science	20 (50)	18 (45)	2 (5)	9 (31)	13 (44.8)	7 (24.1)
A study loses its value if it is published only as a preprint	14 (35)	12 (30)	14 (35)	13 (44.8)	9 (31)	7 (24.2)
How much research parameters investigated in a study contribute to the value of the study	17 (42.5)	7 (17.5)	16 (40)	8 (27.5)	13 (44.8)	8 (27.5)
How many citations a study gets contributes to the value of the study	27 (67.5)	12 (30)	1 (2.5)	27 (93.1)	2 (6.9)	0 (0)
Which journal a study is published in contributes to the value of the study	29 (72.5)	8 (20)	3 (7.5)	25 (86.2)	3 (10.3)	1 (3.4)

[a]All values are n (%).

Subgroup Analysis by Academic Discipline

Participants were also grouped by academic department into basic sciences (n=24) and clinical sciences (n=45). The mean ages were 37.63 and 40.78 years, respectively. Preprint test scores and past use were similar across both groups. However, future intentions diverged: 29.2% (7/24) of basic science participants considered future preprint use, compared to 17.8% (8/45) in the clinical sciences group (data not shown). Additionally, 37.8% (17/45) of clinical faculty reported not considering preprint use in the future, suggesting a more cautious stance (data not shown). Attitudinal data further revealed that clinical scientists were more goal-oriented and focused on applicability in practice, whereas basic scientists placed more emphasis on the breadth of research parameters and openness to publication reform. Notably, more basic science participants believed that studies published only as preprints may lack value (n=14/24, 58.4%) compared to clinical scientists (n=13/45, 28.9%). These findings are visualized in Table 3.

Table . Attitudes and perceptions toward preprints by academic discipline (N=69).[a]

Questions	Basic sciences (n=24)			Clinical sciences (n=45)		
	Agree	Neither agree nor disagree	Disagree	Agree	Neither agree nor disagree	Disagree
Preprints contribute scientific development	16 (66.6)	8 (33.4)	0 (0)	26 (57.8)	15 (33.3)	4 (8.9)
The way preprints change the peer-review and editorial process is favorable for the future of science	13 (54.2)	10 (41.7)	1 (4.2)	16 (35.5)	21 (46.7)	8 (17.8)
A study loses its value of it is published only as a preprint	14 (58.4)	4 (16.7)	6 (25)	13 (28.9)	17 (37.8)	4 (8.9)
How much parameters investigated in a study contributes to the value of study	13 (54.2)	7 (29.2)	4 (16.6)	12 (26.7)	13 (28.9)	18 (44.4)
How much citation gets a study contributes to the value of study	17 (70.9)	6 (25)	1 (4.2)	37 (82.2)	8 (17.8)	0 (0)
Which journal a study published on contributes to the value of study	16 (66.6)	7 (29.2)	1 (4.2)	38 (84.4)	4 (8.9)	3 (6.6)

[a]All values are n (%).

Barriers to Preprint Adoption

Understanding why academics and editors hesitate to engage with preprints is critical for developing targeted interventions. In this study, qualitative responses from survey participants were analyzed thematically and open-ended answers from editors were summarized descriptively to uncover common concerns.

Perceived Barriers Among Medical Academics (Survey Respondents)

Thematic analysis of open-ended responses revealed common concerns including fear of plagiarism or idea theft, lack of academic recognition, and insufficient knowledge about preprints. These barriers are summarized in Table 4.

Table . Common barriers to preprint use identified in participant comments (N=7).

Concern	Frequency (mentions)
Fear of plagiarism or idea theft	7
Preprints not valued in promotion	5
Lack of knowledge about preprints	6
Concerns about ethics or credibility	5
Journal policy restrictions	4

Participants were asked an open-ended question regarding their personal concerns or hesitations about using preprints. Thematic analysis revealed several recurring barriers:

- Fear of plagiarism or idea theft: A frequently mentioned concern was the potential for unreviewed ideas to be copied or republished without attribution. This concern appeared especially prominent among early-career researchers.
- Preprints not valued in promotion: Several participants indicated that preprints are not acknowledged in institutional promotion or academic evaluation processes. As a result, preprints were seen as a risky or unrewarding form of dissemination.
- Lack of knowledge about preprints: Many respondents were unsure about how to submit preprints, what platforms were

reputable, or how preprints interact with formal journal submissions.
- Concerns about ethics or credibility: Some participants questioned whether preprints, by not undergoing peer review, could contribute to the spread of low-quality or misleading research.
- Journal policy restrictions: A few respondents mentioned that they avoided preprints because they believed many journals would reject submissions previously shared as preprints, even if that was not explicitly stated.

These findings suggest that barriers are shaped by both institutional norms and practical uncertainties.

Editorial Perspectives From Turkish Biomedical Journals

Responses from 7 journal editors in Türkiye revealed a spectrum of attitudes toward preprints, ranging from cautious support to open opposition.

One editor expressed support for preprints as a tool to promote transparency and protect authorship in a landscape where idea theft is perceived to be common. Some other editors described preprints as unnecessary for their journal, noting that they already publish accepted articles promptly and that their infrastructure does not support additional processing. Another editor viewed preprints skeptically due to the possibility of their misuse and the ethical complexity of assigning multiple DOIs to similar content. Another editorial opinion highlighted concerns over duplicate publication and the risk that preprints with assigned DOIs might be flagged as plagiarism in similarity checks. Despite these concerns, one editor predicted that preprints would become more widely accepted in the future, potentially reshaping the landscape of academic publishing in Türkiye.

Overall, the responses reflected uncertainty, infrastructural limitations, and a lack of standardization—factors that likely influence journal-level policies and affect how academics perceive the safety and legitimacy of preprint publishing.

Discussion

Our findings reveal a notable gap in both awareness and practical engagement with preprints among medical academics at Marmara University School of Medicine. Despite the global momentum toward open science and rapid communication [14], many respondents exhibited limited familiarity with the concept, and a majority expressed hesitance or skepticism about its use. A recent global survey found that approximately 10% of medical and health sciences researchers in the United States were unfamiliar with preprints, and around 40% had never posted one [15]. In contrast, one-fifth of our participants had never heard of preprints, and four-fifths had never submitted one, highlighting a relatively greater degree of unawareness and hesitancy in our sample.

Misconceptions regarding peer review, duplicate publication, and ethical validity were widespread among respondents, underscoring the need for targeted education and clearer guidance. Maggio and Fleerackers [16] have proposed formally integrating preprint education into health professions curricula to improve understanding and normalize early dissemination practices. Moreover, research shows that students and early career researchers often struggle to distinguish preprints from peer-reviewed journal articles, reflecting a significant knowledge gap [17]. Consistent with these findings, our participants under the age of 40 demonstrated lower preprint knowledge, emphasizing the urgency of tailored interventions for this group.

Journal editors' responses reflected a similar ambivalence. While some recognized the potential of preprints for visibility and transparency, others raised concerns about incompatibility with traditional editorial workflows, ethical ambiguity, and the potential for misuse. Despite such skepticism, studies have shown that peer-reviewed articles that were first posted as preprints tend to receive more citations and broader attention [18], suggesting tangible benefits to early sharing.

Our complementary policy analysis of 118 Turkish biomedical journals further highlights the cultural and structural barriers impeding broader preprint adoption. Only a small proportion of journals explicitly permitted preprints, while the vast majority either prohibited them or failed to mention them altogether. This lack of clear guidance stands in contrast to many high-ranking international clinical journals, which now explicitly allow or even encourage the submission of manuscripts previously posted as preprints [7]. The absence of formal policies among Turkish journals likely contributes to hesitation among authors, reinforcing an environment of academic conservatism and uncertainty—a trend previously noted in the literature [9].

However, our longitudinal review of journal policies between February 2024 and April 2025 suggests a slow but positive shift in this landscape. During this period, 6 additional journals formally adopted policies allowing preprints, while only 1 new prohibition was identified. Although incremental, this trend points toward a growing acceptance and normalization of preprint use within the Turkish biomedical publishing ecosystem. As journal policies become more explicit and aligned with global standards, uncertainty among researchers may diminish, potentially encouraging wider adoption of preprints in academic practice [19].

Additionally, variability in APC policies—particularly the prevalence of "free" and "case-by-case" models—may influence authors' motivations. While one of the advantages of preprints is their cost-free accessibility, this incentive may be undercut if authors are already publishing in fee-free journals or are unaware of preprint benefits [20]. Financial considerations, combined with policy ambiguity, could thus create a disincentive for broader adoption.

Our subgroup analyses revealed how age and academic discipline influence perceptions of preprints. Older academics (≥40 y) were more likely to have published on preprint servers and scored higher on objective knowledge measures but also displayed more conservative attitudes, valuing journal prestige and citation counts. Younger academics (<40 y), on the other hand, were more pessimistic about the future potential of preprints, despite their limited experience. These generational

differences may reflect an evolving academic culture that increasingly values transparency, speed, and accessibility.

This interpretation aligns with the findings of Fraser et al [20], who noted that junior researchers often use preprints to increase the visibility of their work, whereas senior researchers are more motivated by competitive concerns such as staking a priority claim [7].

An interesting finding is that, even though younger academics display reformist attitudes toward the conventional publishing process, they show less openness toward the use of preprints in future. This could be due to their relative inexperience or misconceptions regarding preprints, which can be validated by their lower preprint test score.

We also observed differences between basic and clinical science faculty members. While overall knowledge levels were similar, clinical faculty were more hesitant about preprint use. This may be due to the conservative nature of clinical research, which is closely tied to patient safety, regulatory compliance, and reliance on peer-reviewed evidence. As previously described, clinical medicine tends to be more cautious regarding the role of preprints in academic communication [21].

It is important to acknowledge that this study was conducted at a single academic institution in İstanbul. As such, our findings represent a localized snapshot and should be interpreted with caution. Additionally, while the survey captured a range of perspectives, the response rate was modest, and some questions had incomplete responses. The number of editor responses was also limited, restricting the depth of qualitative analysis. Finally, the policy review was limited to publicly available information, which may not fully reflect internal editorial practices or unpublished updates. Broader, multicenter studies will be necessary to determine whether these patterns hold across other regions and institutions in Türkiye.

To enable broader and more equitable adoption of preprints in the country, we recommend the following:

- Integration of preprints into academic evaluation and promotion criteria
- Development of clear and accessible editorial policies that explicitly state preprint compatibility
- Cross-disciplinary and interinstitutional dialogue to address differing perceptions of scientific quality and publishing ethics
- Encouragement of national or regional preprint platforms tailored to local academic contexts

Bridging both the awareness and policy gaps—and accommodating the diversity of views within academic institutions—is essential for aligning Türkiye's scientific publishing culture with international standards.

This study identified a significant awareness gap and hesitancy toward preprints among medical academics at a Turkish university, driven by both individual misconceptions and a lack of clear, permissive policies from local biomedical journals. Despite a slow trend toward greater journal acceptance, attitudes remain divided by age and discipline. To foster a more open and timely research ecosystem, these findings underscore the need for strategic interventions, including targeted education to correct misconceptions, the establishment of transparent journal policies, and formal institutional recognition of preprints.

Acknowledgments

We would like to thank Burak Kızılca for the assistance provided during the collection of journal policies. All graphs were created in BioRender [22].

Funding

This study did not receive any funding.

Authors' Contributions

Conceptualization, methodology, writing—original draft preparation, writing—review and editing: MS, BK, and ZA. All authors have read and agreed to the published version of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Survey form.
[PDF File, 1810 KB - xmed_v7i1e78139_app1.pdf]

Multimedia Appendix 2
Preprint policies of Turkish biomedical journals.
[XLSX File, 32 KB - xmed_v7i1e78139_app2.xlsx]

Multimedia Appendix 3

Template email sent to editors-in-chief.
[PDF File, 69 KB - xmed_v7i1e78139_app3.pdf]

Checklist 1
CHERRIES checklist.
[PDF File, 164 KB - xmed_v7i1e78139_app4.pdf]

References

1. Smart P. The evolution, benefits, and challenges of preprints and their interaction with journals. Sci Ed 2022 Feb 1;9(1):79-84. [doi: 10.6087/kcse.269]
2. Funk K, Zayas-Cabán T, Beck J. Phase 1 of the National Institutes of Health preprint pilot: testing the viability of making preprints discoverable in PubMed Central and PubMed. bioRxiv. Preprint posted online on Dec 13, 2022. [doi: 10.1101/2022.12.12.520156]
3. Soderberg CK, Errington TM, Nosek BA. Credibility of preprints: an interdisciplinary survey of researchers. R Soc Open Sci 2020 Oct;7(10):201520. [doi: 10.1098/rsos.201520] [Medline: 33204484]
4. Klebel T, Reichmann S, Polka J, et al. Peer review and preprint policies are unclear at most major journals. PLoS One 2020;15(10):e0239518. [doi: 10.1371/journal.pone.0239518] [Medline: 33085678]
5. Choi YJ, Choi HW, Kim S. Preprint acceptance policies of Asian academic society journals in 2020. Sci Ed 2021 Feb 1;8(1):10-17. [doi: 10.6087/kcse.224]
6. Teixeira da Silva JA, Dobránszki J. Preprint policies among 14 academic publishers. The Journal of Academic Librarianship 2019 Mar;45(2):162-170. [doi: 10.1016/j.acalib.2019.02.009]
7. Massey DS, Opare MA, Wallach JD, Ross JS, Krumholz HM. Assessment of preprint policies of top-ranked clinical journals. JAMA Netw Open 2020 Jul 1;3(7):e2011127. [doi: 10.1001/jamanetworkopen.2020.11127] [Medline: 32697320]
8. Sarabipour S, Debat HJ, Emmott E, Burgess SJ, Schwessinger B, Hensel Z. On the value of preprints: an early career researcher perspective. PLoS Biol 2019 Feb;17(2):e3000151. [doi: 10.1371/journal.pbio.3000151] [Medline: 30789895]
9. Ng JY, Chow V, Santoro LJ, et al. An international, cross-sectional survey of preprint attitudes among biomedical researchers. F1000Res. 2024 Nov 4 p. 6. [doi: 10.12688/f1000research.143013.1]
10. Baždarić K, Vrkić I, Arh E, et al. Attitudes and practices of open data, preprinting, and peer-review-a cross sectional study on Croatian scientists. PLoS One 2021;16(6):e0244529. [doi: 10.1371/journal.pone.0244529] [Medline: 34153041]
11. Gutam S, Das S. An overview of publication patterns in India's agricultural research community: journals, open access, and preprints. Res Square. Preprint posted online on May 3, 2023. [doi: 10.21203/rs.3.rs-2565275/v3]
12. Pathak K, Marwaha JS, Chen HW, Krumholz HM, Matthews JB. Open science practices in research published in surgical journals: a cross-sectional study. medRxiv. 2023 May 3 p. 2023.05.02.23289357. [doi: 10.1101/2023.05.02.23289357] [Medline: 37205325]
13. Eysenbach G. Improving the quality of Web surveys: the Checklist for Reporting Results of Internet E-Surveys (CHERRIES). J Med Internet Res 2004 Sep 29;6(3):e34. [doi: 10.2196/jmir.6.3.e34] [Medline: 15471760]
14. Bertram MG, Sundin J, Roche DG, Sánchez-Tójar A, Thoré ESJ, Brodin T. Open science. Curr Biol 2023 Aug 7;33(15):R792-R797. [doi: 10.1016/j.cub.2023.05.036] [Medline: 37552940]
15. Ni R, Waltman L. To preprint or not to preprint: a global researcher survey. Asso for Info Science & Tech 2024 Jun;75(6):749-766. [doi: 10.1002/asi.24880]
16. Maggio LA, Fleerackers A. Preprints in health professions education: raising awareness and shifting culture. Acad Med 2023 Jan 1;98(1):17-20. [doi: 10.1097/ACM.0000000000005001] [Medline: 36576764]
17. Cataldo TT, Faniel IM, Buhler AG, Brannon B, Connaway LS, Putnam S. Students' perceptions of preprints discovered in Google: a window into recognition and evaluation. CRL 2023;84(1):137-156. [doi: 10.5860/crl.84.1.137]
18. Fu DY, Hughey JJ. Releasing a preprint is associated with more attention and citations for the peer-reviewed article. Elife 2019 Dec 6;8:e52646. [doi: 10.7554/eLife.52646] [Medline: 31808742]
19. Springer Nature journals unify their policy to encourage preprint sharing. Nature New Biol 2019 May;569(7756):307-307. [doi: 10.1038/d41586-019-01493-z] [Medline: 31092948]
20. Fraser N, Mayr P, Peters I. Motivations, concerns and selection biases when posting preprints: a survey of bioRxiv authors. PLoS One 2022;17(11):e0274441. [doi: 10.1371/journal.pone.0274441] [Medline: 36327267]
21. Blatch-Jones AJ, Recio Saucedo A, Giddins B. The use and acceptability of preprints in health and social care settings: a scoping review. PLoS ONE 2023;18(9):e0291627. [doi: 10.1371/journal.pone.0291627] [Medline: 37713422]
22. BioRender. BioRender. URL: https://biorender.com/ [accessed 2025-07-28]

Abbreviations

JCI: Journal Citation Index

Edited by S Amal; submitted 27.May.2025; peer-reviewed by K Ide, L Waltman; revised version received 31.Jul.2025; accepted 30.Oct.2025; published 17.Apr.2026.

Please cite as:
Sevim M, Karamese B, Alparslan Z
Awareness, Experiences, and Attitudes Toward Preprints Among Medical Academics: Convergent Mixed Methods Study
JMIRx Med 2026;7:e78139
URL: https://xmed.jmir.org/2026/1/e78139
doi:10.2196/78139

Associations Between IT Job Stressors and Anxiety, Depression, and Stress: Cross-Sectional Study

Edlin Garcia Colato[1], MPH, PhD; Nianjun Liu[2], PhD; Angela Chow[3], PhD; Catherine M Sherwood-Laughlin[3], HSD, MPH; Jonathan T Macy[3], MPH, PhD

[1]Department of Health and Wellness Design, School of Public Health, Indiana University Bloomington, 1025 E 7th St, Bloomington, IN, United States
[2]Department of Epidemiology and Biostatistics, School of Public Health, Indiana University Bloomington, Bloomington, IN, United States
[3]Department of Applied Health Science, School of Public Health, Indiana University Bloomington, Bloomington, IN, United States

Corresponding Author:
Edlin Garcia Colato, MPH, PhD
Department of Health and Wellness Design, School of Public Health, Indiana University Bloomington, 1025 E 7th St, Bloomington, IN, United States

Related Articles:

Companion article: https://www.medrxiv.org/content/10.1101/2025.02.24.25322785v1

Companion article: https://med.jmirx.org/2026/1/e91383

Companion article: https://med.jmirx.org/2026/1/e91437

Abstract

Background: The IT sector is growing and encompasses all professions, from leisure and recreation to hospitals and emergency response groups. IT professionals are experiencing increased threats (eg, ransomware attacks), but little is known about the relationship between these IT profession–specific stressors and the professionals' mental health.

Objective: This study aimed to (1) estimate the associations between IT profession–specific stressors and anxiety, depression, and stress, and (2) examine the role of mental health literacy (MHL) as a mediator of the relationship between depression, anxiety, stress, and help-seeking.

Methods: Between February and May 2023, IT professionals working in the United States were surveyed online. Participants (n=357) reported demographic characteristics, MHL, mental health symptoms, and help-seeking intentions with the following scales: Mental Health Literacy in the Workplace (MHL-W), Center for Epidemiological Studies Depression-10 (CESD-10), Generalized Anxiety Disorder-7 (GAD-7), Perceived Stress Scale-10 (PSS-10), and the Mental Help Seeking Intention Scale (MHSIS). Descriptive statistics, regression models, and mediation analyses were conducted for CESD-10, GAD-7, and PSS-10.

Results: Respondents who had experienced ransomware attacks in the past year reported significantly higher symptoms of depression (odds ratio [OR] 1.85, 95% CI 1.07-3.22; *P*=.03). Past-year exposure to balancing security and usability was associated with lower odds of reported anxiety (OR 0.48, 95% CI 0.28-0.82; *P*=.008). Having made critical technology decisions with limited information in the past year was associated with higher perceived stress by 2.02 points on the PSS-10 scale (SE 0.84, 95% CI 0.37-3.66; *P*=.02), and working with limited resources in the past year increased perceived stress by 1.70 points (SE 0.84, 95% CI 0.04-3.35; *P*=.04) after adjusting for the covariates. MHL was found to partially mediate the relationship between depression and help-seeking, but not between anxiety or stress and help-seeking.

Conclusions: These findings provide insight into the workplace stressors that pose a greater psychological health risk for IT professionals. These results emphasize the important role of MHL in helping facilitate the connection between depressive symptoms and help-seeking.

(JMIRx Med 2026;7:e73211) doi:10.2196/73211

KEYWORDS

survey; association; occupational health; mental health; stressors; IT; IT professionals; United States; workplace; depression; anxiety; stress; help-seeking; health literacy

Introduction

Established in the 1950s, IT is defined as the "use of computer systems or devices to access information" for both business and personal operations, such as "storing, retrieving, accessing or manipulating information" [1]. The efforts of IT workers to maintain business devices can sometimes involve high-stress exposures such as viewing illicit content and mitigating ransomware attacks [2]. Technologies have become persistent targets for hackers and cybercriminals who seek vulnerabilities in networks [3]. Ensuring the safekeeping of technology means that some IT professionals are in a constant state of high alert. The high levels of stress experienced by IT workers are reflected in the findings from a recent assessment of challenges and stressors in the information security sector. According to the 2022 - 2023 Chartered Institute of Information Security report, 22% of 302 UK respondents reported working more than 48 hours per week, with 8% working more than 55 hours weekly [4]. Furthermore, 32% reported they were kept awake by worries of a potential cyber-attack on their organization, up from 22% in 2022 [5]. Over two-thirds of the survey respondents believed there would be an increase in the frequency and impact of ransomware attacks [5].

The existing literature on mental health in IT has largely focused on IT professionals working in Asia [6,7] and Europe [5,8]. Following a ransomware attack, Northwave, a security company based in the Netherlands, conducted a study with 21 members of its own computer emergency response team employees and found that mental health can be significantly affected by ransomware attacks [9]. While the Northwave study suggested an elevated risk of mental health issues among IT professionals, another study using the UK Biobank cohort study did not yield similar findings [10]. This was the first UK-based longitudinal study (running from 2006 to 2010); it compared the incidence of anxiety and depression between IT and non-IT employees aged 40 years or older found that IT professionals had a reduced risk of anxiety and depression compared to their non-IT counterparts [10]. To the best of our knowledge, no such report exists for the IT sector in the United States.

Considering the IT professionals who report experiencing symptoms of depression, anxiety, and stress, an important next step is to determine whether these individuals possess a firm knowledge base on mental health and whether they intend to seek help. Mental health literacy (MHL) refers to the knowledge and ability to recognize and identify symptoms related to mental health for preventing mental illness as well as maintaining and promoting mental health [11]. Previous research has found a positive correlation between MHL and mental health attitudes [12,13], and young adults with more favorable attitudes toward mental health services are more likely to seek help [14].

Past studies have not assessed MHL, mental health attitudes, or intention to seek help among IT workers in the United States. Additionally, previous studies have not considered important factors such as exposure to illicit content, ransomware attacks, hacking, and takedowns, which may contribute to the elevated symptoms of depression, anxiety, and stress experienced by IT professionals. A recently published occasional paper from the United Kingdom examined first-order harms of ransomware on staff and identified the stress reported by incident responders following the incident [15]. First-order harms are "harms to any organization and their staff directly targeted by a ransomware incident" [4]. There is an urgent need for evidence-based resources and information concerning mental health in the IT workforce community, a domain that remains under-studied, especially in the US context. Therefore, this study aimed to (1) test the relationship between IT profession–specific stressors and anxiety, depression, and stress and (2) analyze the role of MHL as a mediator for the 3 selected mental health conditions (anxiety, depression, and stress) and help-seeking behaviors.

Methods

Participants and Data Collection

Data for this study were collected between February and May 2023 via an online survey using Qualtrics [16]. The online open survey was developed with previously validated scales described in detail in the Measures section below. For this convenience sample, 2336 participants were identified via known contacts, SurveyCircle [17], and Prolific [18]. The three eligibility criteria for the survey were (1) being aged at least 18 years, (2) working in the IT sector in the United States, and (3) having at least 12 months of any IT work experience. Electronic written voluntary consent was recorded for all survey respondents at the start of the survey. Duplicate entries were avoided by limiting access to the survey to a single attempt. Of the original 483 responses recorded, 23 observations were excluded because of either discontinuing the survey prior to providing consent, being ineligible, or being a bot entry. Review of the data showed no evidence of conspicuous response behavior. Nevertheless, outliers in average completion time of the survey, such as those showing that the survey was completed in only a few minutes, were excluded. A total of 388 (84.3%) of the remaining 460 participants who provided consent and whose responses were determined to be valid completed the survey. The final sample following complete case handling of missing data is described below in the Data Analysis section.

Ethical Considerations

The consent form included pertinent information, such as the estimated duration of the survey, the fact that personally identifiable information would not be collected, and the purpose of the study. After respondents completed the 10-to-15-minute survey, they were given the option to take part in a drawing for one of four US $75 Target e-gift cards. To ensure anonymity in the survey responses, respondents were guided to a separate Qualtrics survey where they could provide an email address for the random drawing. This study received approval as an exempt study by the Indiana University Bloomington Human Subjects and Institutional Review Board (protocol 18281).

Measures

Assessment of Mental Health Status

Symptoms of depression were assessed with the 10-item Center for Epidemiologic Studies Depression (CESD-10) questionnaire [19,20]. The CESD-10 helps identify individuals at risk of developing clinical depression and has been previously validated

against the original 20-item CESD questionnaire [21] designed for screening the general population [19,20,22]. CESD-10 scores range between 0 and 30; scores below 10 are recoded to 0 ("no significant symptoms of depression") and scores 10 and above are recoded to 1 ("significant symptoms of depression") creating a binary variable for depression.

Meanwhile, the 7-item Generalized Anxiety Disorder Scale (GAD-7) was used to screen for symptoms of anxiety among the respondents [23]. Responses to the 7 questions were summed for a final score ranging from 0 to 21. For descriptive purposes, 3 cutoff points were used (5, 10, and 15; 0‐4=minimal anxiety, 5‐9=mild anxiety, 10‐14=moderate anxiety, and 15‐21=severe anxiety) to show the different categorical levels of severity. However, the optimal cutoff score for screening anxiety via the GAD-7 scale is 10 [23,24]. Therefore, scores 10 and above were coded as 1 ("yes") for anxiety and below 10 were coded as 0 ("no/minimal") for level of anxiety.

Unlike the anxiety and depression scales, the 10-item Perceived Stress Scale (PSS-10) does not translate to clinical significance; therefore, categorizing scores into groups is done only for descriptive purposes, and for the analyses the scores were included as a continuous variable [25]. Final scores ranged between 0 and 40. For descriptive purposes, scores 0 to 13 were coded as "low stress," 14 to 26 as "moderate stress," and scores 27 to 40 as "high stress." Increased perceived stress is reflected by higher scores.

IT Profession–Specific Stressors

For research question 1, we identified 12 past-year IT profession–specific stressors based on the feedback solicited from 2 IT professionals with a combined 35 years of IT experience. IT experts were selected based on predefined criteria, including their professional qualifications and practical experience. Interviews were conducted with the IT experts, allowing for the development of a comprehensive list of stressors. The list of the 12 stressors curated by the IT professionals was as follows: (1) ransomware attacks, (2) exposure to illicit content, (3) takedowns, (4) handling sensitive data and cybersecurity threats, (5) making critical technology decisions with limited information, 6) adapting to rapid changes in technology and business requirements, (7) pressure to solve complex technical issues, (8) the constant need to stay up to date with technology, (9) dealing with unexpected system failures and outages, (10) dealing with leadership that does not wish to invest in or be inconvenienced by cybersecurity initiatives, (11) balancing security and usability, and (12) working with limited resources (eg, budget and personnel).

Respondents were asked a single question: "Which of the following on-the-job stressors have you experienced in the past year?" and selected all that applied from the list of stressors provided. The 12 stressors were measured individually as binary variables (yes/no), and a separate count variable was created to identify the total number of the 12 stressors experienced by each respondent (scores ranged from 0 to 12).

Mental Health Help-Seeking Intentions

For research question 2, the primary outcome was mental health help-seeking intentions, which were measured using the Mental Help Seeking Intention Scale (MHSIS) [26]. Final MHSIS scores ranged between 1 and 7. Higher scores indicate greater intention to seek help from a mental health professional [26].

Mental Health Literacy in the Workplace

The Mental Health Literacy Tool for the Workplace (MHL-W) comprises 4 vignettes depicting a hypothetical coworker's behavior in the workplace; each vignette has 4 questions measuring 4 distinct MHL concepts [27]. The following 4 variables were rated on a scale from 1 (very low) to 5 (very high): (1) level of knowledge in being able to recognize a specific disorder ("What might be happening with [him/her]),"(2) level of knowledge and beliefs regarding risk factors and prevention ("How you could prevent the situation from becoming worse"), (3) level of knowledge and attitudes about help-seeking ("What you should say or do in the situation"), and (4) level of knowledge and beliefs regarding interventions ("Resources or services that might be helpful"). The values for all 16 questions were summed up for a final MHL-W score, with possible scores ranging from 16 to 80. Higher scores indicate a greater self-reported knowledge of mental health.

Demographics

Demographic data were collected on the following characteristics: age, race (Black or African American, American Indian or Alaska Native, Asian or Asian American, Native Hawaiian or other Pacific Islander, White, and other), ethnicity, sex at birth (female vs male), relationship status, education, and income-level groups. Data regarding health-related characteristics included questions about health insurance and mental health history.

Covariates

Covariates included age, sex, race, ethnicity, education, income, health insurance, and mental health history.

Data Analysis

Descriptive statistics, means and SD for continuous variables, and frequencies and percentages for categorical variables were computed by respondent sex at birth. A 2-sided *P* value of less than .05 was considered significant. A complete case analysis was performed, excluding participants with any missing data for key variables. Of the 388 participants, 17 (4%) were missing data for age, 1 (<1%) for race, 4 (1%) for ethnicity, 1 (<1%) for education, 2 (<1%) for health insurance, 5 (1%) for CESD-10, 1 (<1%) for GAD-7, 13 (3%) for PSS-10, 1 (<1%) for MHSIS, and 10 (3%) for MHL-W. A total of 357 participants had complete data on all key variables.

The following models were used to address the first hypothesis and identify which IT profession–specific stressors were associated with anxiety, depression, and stress: (1) unadjusted and adjusted logistic regression models to examine the relationship between each of the 12 IT-profession specific stressors and depression (CESD-10), (2) unadjusted and predictor-adjusted logistic regression models to test for the relationship between each of the 12 IT-profession specific stressors and anxiety (GAD-7), and (3) unadjusted and predictor-adjusted linear regression models to test for the

association between each of the 12 IT-specific stressors and stress (PSS-10).

A confirmatory factor analysis (CFA) was conducted for the MHL-W because the workplace questionnaire had previously been used only in a single workplace setting in health care. The CFA aimed to test whether each of the 4 questions per vignette loaded onto the intended corresponding MHL constructs.

Mediation Analysis

Mediation analysis was conducted using the *medsem* package in Stata (StataSE version 17.0; StataCorp). Outputs for the structural equation modeling for the mediation analysis included a modified version of the Sobel test for assessing indirect effects [28], as well as the Monte Carlo resampling approach for assessing indirect effects [29]. Descriptive statistics, a CFA, mediation, and all other analyses were conducted using Stata.

Results

Participants

Figure 1 shows a flowchart of participation, and Table 1 shows participant characteristics. The majority of the 357 respondents self-identified as male (n=264; 73.9%), were White (n=281; 78.7%), were of non-Hispanic/Latinx ethnicity (n=323; 90.5%), had earned a bachelor's degree (n=181; 50.7%), and earned US $75,000 or more a year (n=213; 59.7%).

Figure 1. Flow diagram of the study sample.

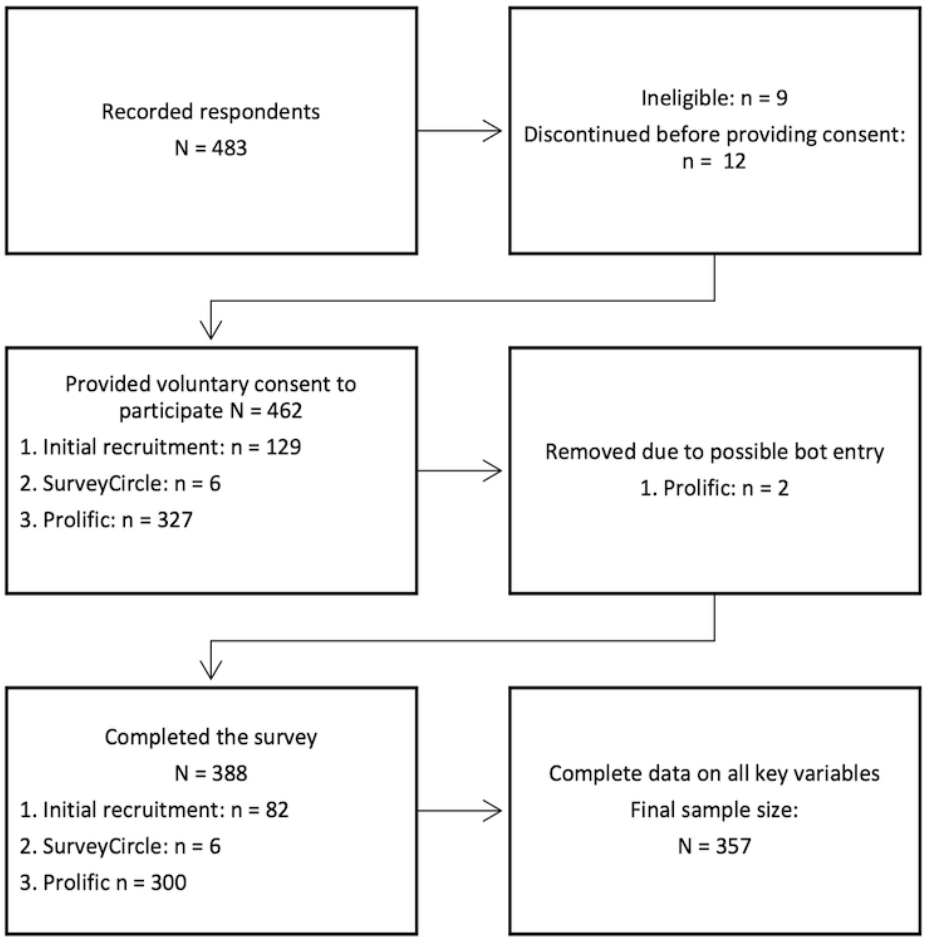


Table . Participant characteristics by sex (n=357).

		Female (n=93; 26.1%)	Male (n=264; 73.9%)	Total
Age (years), median (SD)		42 (11.4)	39 (10.7)	39 (10.9)
Race, n (%)				
	African American or Black	4 (4.3)	12 (4.5)	16 (4.5)
	American Indian or Alaskan Native	0 (0.0)	3 (1.1)	3 (0.8)
	Asian or Asian American	12 (12.9)	31 (11.7)	43 (12.0)
	White	74 (79.6)	207 (78.4)	281 (78.7)
	Other	1 (1.1)	3 (1.1)	4 (1.1)
	Two or more selected	2 (2.2)	8 (3.0)	10 (2.8)
Ethnicity, n (%)				
	Non-Hispanic	82 (88.2)	241 (91.3)	323 (90.5)
	Hispanic	11 (11.8)	23 (8.7)	34 (9.5)
Education, n (%)				
	Less than a bachelor's degree	22 (237)	83 (31.4)	105 (29.4)
	Bachelor's degree	43 (46.2)	138 (52.3)	181 (50.7)
	Graduate's degree	28 (30.1)	43 (16.3)	71 (19.9)
Health insurance type, n (%)				
	Employer provided	61 (65.6)	176 (66.7)	237 (66.4)
	Government provided	27 (29.0)	78 (29.5)	105 (29.4)
	None	5 (5.4)	10 (3.8)	15 (4.2)
Income group (US $), n (%)				
	Less than 50,000	18 (19.4)	52 (19.7)	70 (19.6)
	50,000-74,999	20 (21.5)	54 (20.5)	74 (20.7)
	75,000-99,999	23 (24.7)	48 (18.2)	71 (19.9)
	≥100,000	32 (34.4)	110 (41.7)	142 (39.8)
Mental health history, n (%)				
	Yes	42 (45.2)	57 (21.6)	99 (27.7)
	No	51 (54.8)	207 (78.4)	258 (72.3)
CESD-10[a] score, n (%)				
	No/low	50 (53.8)	181 (68.6)	231 (64.7)
	Moderate/severe	43 (46.2)	83 (31.4)	126 (35.3)
GAD-7[b] score, n (%)				
	No/low	72 (77.4)	215 (81.4)	287 (80.4)
	Moderate/severe	21 (22.6)	49 (18.6)	70 (19.6)
MHSIS[c] score, mean (SD)		5.33 (1.5)	5.11 (1.7)	5.17 (1.7)
MHL-W[d] score, mean (SD)		47.95 (11.9)	46.46 (10.9)	46.85 (11.2)
PSS-10[e] score, mean (SD)		17.15 (7.9)	14.30 (8.1)	15.04 (8.1)

[a]CESD-10: 10-item Center for Epidemiologic Studies Depression Scale.

[b]GAD-7: 7-item Generalized Anxiety Disorder.

[c]MHSIS: Mental Help Seeking Intention Scale.

[d]MHL-W: Mental Health Literacy at the Workplace.

[e]PSS-10: 10-item Perceived Stress Scale.

Outcome Data

The majority of participants (231/357, 64.7%) were below the minimum cutoff score of 10 for depression as assessed by the CESD-10. Similarly, 287 of 357 (80.4%) respondents reported having no or low symptoms of anxiety. The mean PSS-10 score was 15.04 (SD 8.1), representing a moderate level of stress. For mental health history, most (258/357, 72.3%) of the respondents reported having no known previous mental health diagnosis. On average, 5 stressors were experienced in this population, as shown in Table S1 in Multimedia Appendix 1. Past-year takedowns were experienced the least (75/357, 21.1%) by the sample, while experiencing pressure to solve complex technical issues was the most common experience (252/357, 70.1%).

Main Results

IT Profession–Specific Stressors and Depression

Prior to controlling for sociodemographic variables, none of the stressors were significantly associated with depression. The respondents who reported past-year exposure to ransomware (odds ratio [OR] 1.85, 95% CI 1.07-3.22; *P*=.03) or working with leadership that did not wish to invest in or be inconvenienced by cybersecurity initiatives (OR 1.75, 95% CI 1.01-2.97; *P*=.048) in the past year were more likely to have significant symptoms of depression, after controlling for the covariates. The total number of stressors experienced in the past year also had no significant association with depressive symptoms, even after controlling for the covariates. None of the remaining IT profession–specific stressors experienced in the past year were individually associated with significant symptoms of depression.

IT Profession–Specific Stressors and Anxiety

Past-year exposure to balancing security and usability was associated with lower odds of reported anxiety (OR 0.48, 95% CI 0.28-0.82; *P*=.008); however, that association did not hold once adjusting for the covariates. All remaining IT stressors had no significant association with anxiety, including after controlling for the covariates.

IT Profession–Specific Stressors and Stress

Controlling for the covariates, respondents who reported making critical technology decisions with limited information in the past year had higher perceived stress by 2.02 points on the PSS-10 scale (SE 0.84, 95% CI 0.37-3.66; *P*=.02). Similarly, respondents who reported working with limited resources (eg, budget and personnel) in the past year had higher perceived stress by 1.70 points on the PSS-10 scale (SE 0.84, 95% CI 0.04-3.35; *P*=.04), after adjusting for the covariates. The remaining stressors did not have a significant relationship with perceived stress.

Mediation Analysis

Mediation analysis was performed to assess the mediating role of MHL-W score in the relationship between depression, anxiety, and stress scores (separately) and help-seeking intentions (MHSIS). The results (shown in Table 2) revealed that depression also significantly predicted MHL-W score (coefficient=–2.236; *P*=.049) and that MHL-W score significantly predicted MHSIS score (coefficient=0.037; *P*<.001). After controlling for MHL-W score, the direct effect of depression on MHSIS score remained significant (coefficient=–0.438; *P*=.01). Per the Sobel test and the Monte Carlo approaches, MHL-W score partially mediated the relationship between depression and help-seeking intentions.

Table . Mediation results for mental health literacy and depression, anxiety, and stress.

Variable and effect type	Coefficient	Z value	*P* value	95% CI
CESD-10[a] → MHL-W[e] → MHSIS[f]				
Total effect (CESD-10 → MHSIS)	–0.524	—[d]	—	—
Direct effects (CESD-10 → MHSIS)	–0.438	–2.53	.01	–0.778 to –0.098
Indirect effects of CESD-10 on MHSIS	–0.086	–1.826	.07	–0.194 to –0.003
GAD-7[b] → MHL-W → MHSIS				
Total effect (GAD-7 → MHSIS)	0.033	—	—	—
Direct effects (GAD-7 → MHSIS)	–0.517	–2.49	.01	–0.924 to –0.110
Indirect effects of GAD-7 on MHSIS	–0.018	–0.31	.75	–0.128 to 0.093
PSS-10[c] → MHL-W → MHSIS				
Total effect (PSS-10→ MHSIS)	–0.048	—	—	—
Direct effects (PSS-10→ MHSIS)	–0.028	–2.83	.005	–0.048 to –0.009
Indirect effects of PSS-10 on MHSIS	–0.005	0.003	.13	–0.011 to 0.001

[a]CESD-10: 10-item Center for Epidemiological Studies Depression.

[b]GAD-7: 7-item Generalized Anxiety Disorder.

[c]PSS-10: 10-item Perceived Stress Scale.

[d]Not applicable.

[e]MHL-W: Mental Health Literacy in the Workplace.

[f]MHSIS: Mental Help Seeking Intention Scale.

On the contrary, the results for mediation analyses for anxiety and stress did not support MHL-W score as a mediator for anxiety or stress and help-seeking intentions. Although there was a significant relationship between MHL-W score and MHSIS score, there was no significant relationship between GAD-7 score or PSS-10 score and MHL-W score.

Discussion

Key Results

This study was conducted to (1) test which IT profession–specific stressors were associated with depression, anxiety, and stress; and (2) examine the impact of depression, anxiety, and stress on help-seeking as mediated by mental health literacy. For the first objective, we found that past exposure to ransomware attacks and working with leadership that did not wish to invest in or were inconvenienced by cybersecurity initiatives were both associated with higher odds of depressive symptoms. While none of the stressors had a significant relationship with anxiety, two stressors did have a significant relationship with perceived stress. Making critical technology decisions with limited information and working with limited resources were both linked to higher perceived stress. Regarding the second objective, we found that mental health literacy in the workplace only partially mediated the relationship between depression and help-seeking intentions, but not between anxiety or stress and help-seeking intentions.

The findings for the first objective align with and complement findings from prior work. For instance, Northwave's report on the psychological impact of a ransomware attack on their employees showed there was significant stress experienced immediately after the incident and throughout the following year [9], while in this study, ransomware was associated with symptoms of depression and not current perceived stress. This might be because the stress instrument used captures general stress over the past month and is not event-specific. However, the depression questionnaire used captured longer-term emotional stress, which develops over time after prolonged exposure to high-stakes incidents (eg, ransomware). Furthermore, having a leadership team that is unsupportive of cybersecurity initiatives, which would help deter ransomware attacks, was also associated with increased depressive symptoms. This adds to the ransomware literature, which is heavily focused on the financial impact on individuals and organizations and the psychosocial costs to the victims, rather than considering the individuals working tirelessly to mitigate the problem [30,31]. Only two stressors had a significant association with increased perceived stress: making critical

decisions with limited information and working with limited resources. Both of these stressors and increased stress align with the job demands–resources model, in which having inadequate resources places strain on the worker, increasing risk of burnout [32,33].

The pressure to solve complex technical issues, the constant need to stay up to date with technology, and the task of dealing with unexpected system failures and outages were not associated with depression, anxiety, or stress. This is likely because these stressors, except for the unexpected system failures and outages, are typical daily job activities and expected tasks and responsibilities for individuals working in the IT sector. Staying up to date with technology and solving complex technical issues are part of the allure of working in technology. These may act as positive stressors, potentially increasing stress while also promoting professional and personal growth, rather than predominantly negative experiences for IT professionals [34]. The influence of positive and welcomed anticipated stressors among IT professionals requires further exploration.

Interestingly, exposure to illicit content was not associated with depression, anxiety, or stress in this study. Previous literature has found that exposure to illicit content (eg, exploitative, violent, and/or abusive content) can have psychologically detrimental effects on children and adolescents [35]. While similar associations have not been extensively documented among adults, Federal Bureau of Investigation officers have developed coping strategies to deal with exposure to illicit content. For officers who are employed in this line of work, the skill of compartmentalization is essential for sustaining their professional role [36]. Future research should examine the type and amount of illicit content exposure to increase our understanding of when and how this content becomes harmful, specifically among adults and as an occupational hazard.

MHL in the workplace continues to be an important area of focus [37]. The mediation results suggest a pathway worth noting between MHL and help-seeking for depression, although the indirect effect is modest. MHL was a partial mediator for depression, but not for anxiety or stress. The lack of mediation found between MHL and anxiety and stress could be due to the measurement timing or the cross-sectional design, which limited the ability to detect indirect effects for both anxiety and stress. Future research using longitudinal designs and alternative mediators could clarify whether these null findings reflect a true absence of mediation or are the result of methodological constraints. Unlike depression and anxiety, stress is not a medical diagnosis, but rather an automatic bodily response that is experienced after certain events. Therefore, it often does not require medical help. Stress is experienced daily and stress levels constantly fluctuate [38]. Chronic stress, on the other hand, can increase the risk for depression and anxiety [39], making it important to note that the PSS-10 instrument used in this study captured past-week perceived stress, not chronic stress.

A 2019 UK industry report by the British Interactive Media Association indicated that 28% of survey respondents experienced past-year anxiety and depression [8], a percentage that is similar to the sample in this study, 28.1% of whom reported ever having been diagnosed with a mental health condition. On the other hand, for current depression and anxiety, measured via the CESD-10 and GAD-7 scores, 29.3% of respondents were assessed as having recent symptoms of clinically significant depression and mild or greater symptoms of anxiety. These numbers are much lower than the 81% reported by the British Interactive Media Association. Differences in these numbers could be due to how the questions were formulated. This study used two validated instruments to capture symptoms, while the UK industry survey asked respondents to self-report past-year (12-month) anxiety and depression experiences.

Unfortunately, the results of this study are not directly comparable to those of the UK Biobank cohort study. It is important to note that the UK Biobank cohort study, in addition to being a longitudinal study with a UK sample, measured depression and anxiety with a single assessment tool, the Patient Health Questionnaire-4, while this US-focused study included the CESD-10 and GAD-7. Nevertheless, this study is the first to assess the relationship between IT profession–specific stressors and depression, anxiety, and stress, aiming to shed light on their associations. It also attempted to uncover the mediating role of mental health literacy between mental health conditions and help-seeking intentions.

Limitations and Generalizability

There are a few limitations that must be noted. First, this study used a cross-sectional design, and although a mediation analysis was conducted, neither direction nor cause and effect of the relationships can be determined. The relationships found are associations and must not be regarded as causal. Furthermore, the instrument, MHL-W, used to ascertain MHL relies on self-reported level of knowledge, which might be overinflated. However, this is the only validated MHL instrument that is specific to the workplace, making it the most suitable for this type of research.

Although the sample resembled the demographic characteristics of the 357 IT professionals overall by sex (n=264, 73.9% male vs n=93, 26.1% female) [40], this sample is slightly overrepresented by White respondents (281/357, 78.7%). The 2019 US Census Bureau estimates high technology industries are comprised of primarily White (68.5%), followed by Asian (14%) IT professionals [41]. Although this sample is not representative of the demographic characteristics of the current IT sector, the goal of this study was to understand the relationship between exposures and outcome regardless of demographic information.

Despite the limitations, the results of this exploratory study are promising in that they provide insight into the types of stressors that pose a greater psychological health risk for IT professionals. Again, the possible high-risk stressors identified in this study include past-year exposure to ransomware attacks, making critical technical decisions with limited information, and dealing with leadership that is not interested in cybersecurity initiatives. These results might also be beneficial to IT leaders when considering what response plans are in place to help mitigate the psychosocial impacts these exposures may have on the IT professionals in their organizations. As has been shown

previously, leadership training in mental health resources impacts employees' willingness and actual use of those resources [42]. Furthermore, the findings highlight the important role of MHL in helping facilitate the connection between experiencing significant symptoms of depression and seeking help to address these symptoms.

Conclusion

As the IT workforce continues to expand throughout many other sectors, thus increasing IT professional opportunities for employment, the lack of comprehensive mental health support and resources in some sectors could have negative consequences for their overall quality of life. Their health and well-being are crucial not only from a worker and an industry-level perspective, but also because of the significance of their job roles. Gaining insight into their mental health needs could facilitate the development of strategies for implementing programs aimed at improving and/or maintaining good mental health.

Before leaders can identify solutions and resources for their teams, we need leaders to be on board with supporting these teams. Given the importance of employee safety, health, and well-being, results from this study can aid IT leaders in identifying situations where IT professionals could be at an elevated risk due to workplace stressors. This awareness would enable them to proactively provide resources and support in a timely fashion. For IT-specific stressors, it is not a matter of if, but rather when, situations such as a ransomware attack will occur. Therefore, having a plan that not only focuses on the attack itself, but the human being behind the computer will be essential for promoting health among IT professionals.

Funding

The authors received no external financial support for the research, authorship, and/or publication of this article.

Data Availability

The datasets generated during and/or analyzed during the current study are available from the corresponding author upon reasonable request.

Authors' Contributions

Conceptualization: EGC
Data curation: EGC
Formal analysis: EGC
Funding acquisition: EGC
Investigation: EGC
Methodology: EGC, NL
Project administration: JTM
Resources: JTM
Supervision: JTM, CS-L, NL, AC
Validation: EGC, NL
Visualization: EGC
Writing—original draft: EGC
Writing—review and editing: EGC, JTM, NL, AC, CS-L

Conflicts of Interest

None declared.

Multimedia Appendix 1
Stressors experienced by the participants.
[DOCX File, 20 KB - xmed_v7i1e73211_app1.docx]

References

1. What is information technology? CompTIA. URL: https://www.comptia.org/content/articles/what-is-information-technology [accessed 2023-01-30]
2. Boehm MA, Lei QM, Lloyd RM, Prichard JR. Depression, anxiety, and tobacco use: overlapping impediments to sleep in a national sample of college students. J Am Coll Health 2016 Oct;64(7):565-574. [doi: 10.1080/07448481.2016.1205073] [Medline: 27347758]
3. Samtani S, Chinn R, Chen H, Nunamaker JF. Exploring emerging hacker assets and key hackers for proactive cyber threat intelligence. J Manage Inf Syst 2017 Oct 2;34(4):1023-1053. [doi: 10.1080/07421222.2017.1394049]
4. The security profession in 2022-23. Chartered Institute of Information Security. URL: https://www.ciisec.org/survey/the-security-profession-in-2022-23/ [accessed 2024-05-05]

5. Our industry survey. Chartered Institute of Information Security. URL: https://www.ciisec.org/White_Papers [accessed 2023-04-27]
6. Lim VKG, Teo TSH. Occupational stress and IT personnel in Singapore: factorial dimensions and differential effects. Int J Inf Manage 1999 Aug;19(4):277-291. [doi: 10.1016/S0268-4012(99)00027-4]
7. Rao JV, Chandraiah K. Occupational stress, mental health and coping among information technology professionals. Indian J Occup Environ Med 2012 Jan;16(1):22-26. [doi: 10.4103/0019-5278.99686] [Medline: 23112503]
8. The voices of our industry: BIMA tech inclusion & diversity report 2019. British Interactive Media Association. URL: https://bima.co.uk/wp-content/uploads/2020/01/BIMA-Tech-Inclusion-and-Diversity-Report-2019.pdf [accessed 2022-10-29]
9. After the crisis comes the blow - the mental impact of ransomware attacks. Northwave. 2022. URL: https://northwave-cybersecurity.com/whitepapers-articles/after-the-crisis-comes-the-blow [accessed 2023-02-15]
10. Lalloo D, Lewsey J, Katikireddi SV, Macdonald EB, Campbell D, Demou E. Comparing anxiety and depression in information technology workers with others in employment: a UK Biobank cohort study. Ann Work Expo Health 2022 Nov 15;66(9):1136-1150. [doi: 10.1093/annweh/wxac061] [Medline: 36029464]
11. Jorm AF, Korten AE, Jacomb PA, Christensen H, Rodgers B, Pollitt P. "Mental health literacy": a survey of the public's ability to recognise mental disorders and their beliefs about the effectiveness of treatment. Med J Aust 1997 Feb 17;166(4):182-186. [doi: 10.5694/j.1326-5377.1997.tb140071.x] [Medline: 9066546]
12. Lee HY, Hwang J, Ball JG, Lee J, Yu Y, Albright DL. Mental health literacy affects mental health attitude: is there a gender difference? Am J Health Behav 2020 May 1;44(3):282-291. [doi: 10.5993/AJHB.44.3.1] [Medline: 32295676]
13. Yeo G, Reich SM, Liaw NA, Chia EYM. The effect of digital mental health literacy interventions on mental health: systematic review and meta-analysis. J Med Internet Res 2024 Feb 29;26:e51268. [doi: 10.2196/51268] [Medline: 38421687]
14. Zorrilla MM, Modeste N, Gleason PC, Sealy DA, Banta JE, Trieu SL. Depression and help-seeking intention among young adults: the theory of planned behavior. Am J Health Educ 2019 Jul 4;50(4):236-244. [doi: 10.1080/19325037.2019.1616014]
15. Jamie MacColl PH, Mott G, James Sullivan J, Sarah P. The scourge of ransomware: victim insights on harms to individuals, organisations and society. Royal United Services Institute for Defence and Security Studies. 2024. URL: https://static.rusi.org/ransomware-harms-op-january-2024.pdf [accessed 2024-11-01]
16. Qualtrics: university information technology services. Indiana University. 2022. URL: https://uits.iu.edu/qualtrics [accessed 2022-12-01]
17. FAQ – frequently asked questions. SurveyCircle. 2023. URL: https://www.surveycircle.com/en/faq [accessed 2023-02-15]
18. Conduct your best research. Prolific. 2023. URL: https://www.prolific.co/researchers#pricing [accessed 2023-02-20]
19. Radloff LS. The CES-D scale: a self-report depression scale for research in the general population. Appl Psychol Meas 1977;1(3):385-401. [doi: 10.1177/014662167700100306]
20. Vilagut G, Forero CG, Barbaglia G, Alonso J. Screening for depression in the general population with the Center for Epidemiologic Studies Depression (CES-D): a systematic review with meta-analysis. PLoS ONE 2016;11(5):e0155431. [doi: 10.1371/journal.pone.0155431] [Medline: 27182821]
21. Shrout PE, Yager TJ. Reliability and validity of screening scales: effect of reducing scale length. J Clin Epidemiol 1989;42(1):69-78. [doi: 10.1016/0895-4356(89)90027-9] [Medline: 2913189]
22. Perreira KM, Deeb-Sossa N, Harris KM, Bollen K. What are we measuring? An evaluation of the CES-D across race/ethnicity and immigrant generation*. Soc Forces 2005 Jun;83(4):1567-1601. [doi: 10.1353/sof.2005.0077]
23. Spitzer RL, Kroenke K, Williams JBW, Löwe B. A brief measure for assessing generalized anxiety disorder: the GAD-7. Arch Intern Med 2006 May 22;166(10):1092-1097. [doi: 10.1001/archinte.166.10.1092] [Medline: 16717171]
24. Johnson SU, Ulvenes PG, Øktedalen T, Hoffart A. Psychometric properties of the General Anxiety Disorder 7-Item (GAD-7) scale in a heterogeneous psychiatric sample. Front Psychol 2019;10(1713):1713. [doi: 10.3389/fpsyg.2019.01713] [Medline: 31447721]
25. Cohen S, Kamarck T, Mermelstein R. A global measure of perceived stress. J Health Soc Behav 1983 Dec;24(4):385-396. [doi: 10.2307/2136404] [Medline: 6668417]
26. Hammer JH, Spiker DA. Dimensionality, reliability, and predictive evidence of validity for three help-seeking intention instruments: ISCI, GHSQ, and MHSIS. J Couns Psychol 2018 Apr;65(3):394-401. [doi: 10.1037/cou0000256] [Medline: 29672088]
27. Moll S, Zanhour M, Patten SB, Stuart H, MacDermid J. Evaluating mental health literacy in the workplace: development and psychometric properties of a vignette-based tool. J Occup Rehabil 2017 Dec;27(4):601-611. [doi: 10.1007/s10926-017-9695-0] [Medline: 28120136]
28. Baron RM, Kenny DA. The moderator-mediator variable distinction in social psychological research: conceptual, strategic, and statistical considerations. J Pers Soc Psychol 1986 Dec;51(6):1173-1182. [doi: 10.1037//0022-3514.51.6.1173] [Medline: 3806354]
29. Zhao X, Lynch JG, Chen Q. Reconsidering Baron and Kenny: myths and truths about mediation analysis. J Consum Res 2010 Aug;37(2):197-206. [doi: 10.1086/651257]
30. Chen C. The human consequences of ransomware attacks. Information Systems Audit and Control Association. 2022. URL: https://www.isaca.org/resources/isaca-journal/issues/2022/volume-3/the-human-consequences-of-ransomware-attacks [accessed 2026-01-21]

31. van Boven LS, Kusters RWJ, Tin D, et al. Hacking acute care: a qualitative study on the health care impacts of ransomware attacks against hospitals. Ann Emerg Med 2024 Jan;83(1):46-56. [doi: 10.1016/j.annemergmed.2023.04.025] [Medline: 37318433]
32. Demerouti E, Bakker AB, Nachreiner F, Schaufeli WB. The job demands-resources model of burnout. J Appl Psychol 2001 Jun;86(3):499-512. [doi: 10.1037/0021-9010.86.3.499] [Medline: 11419809]
33. Zaza S, Riemenschneider C, Armstrong DJ. The drivers and effects of burnout within an information technology work context: a job demands-resources framework. Inf Technol People 2022 Dec 7;35(7):2288-2313. [doi: 10.1108/ITP-01-2021-0093]
34. Jensen C. What attracts people to IT? The Forecast by Nutanix. 2019. URL: https://www.nutanix.com/theforecastbynutanix/business/what-attracts-people-to-it [accessed 2022-09-07]
35. Meates J. Problematic digital technology use in children and adolescents: impact on physical well-being. Teach Curric 2021;21(1):77-91. [doi: 10.15663/tandc.v21i1.363]
36. Cruz N. FBI Law Enforcement Bulletin. Safeguard spotlight: coping with line-of-duty exposure to child pornography/exploitation materials. 2011. URL: https://leb.fbi.gov/spotlights/safeguard-spotlight-coping-with-line-of-duty-exposure-to-child-pornographyexploitation-materials [accessed 2022-08-05]
37. Lam LT, Lam MKP. A web-based and mobile intervention program using a spaced education approach for workplace mental health literacy: cluster randomized controlled trial. JMIR Ment Health 2024 Apr 23;11:e51791. [doi: 10.2196/51791] [Medline: 38654570]
38. World Health Organization. Stress. 2023. URL: https://www.who.int/news-room/questions-and-answers/item/stress [accessed 2023-02-25]
39. Tafet GE, Bernardini R. Psychoneuroendocrinological links between chronic stress and depression. Prog Neuropsychopharmacol Biol Psychiatry 2003 Sep;27(6):893-903. [doi: 10.1016/S0278-5846(03)00162-3] [Medline: 14499305]
40. White SK. Women in tech statistics: the hard truths of an uphill battle. CIO. 2023. URL: https://www.cio.com/article/201905/women-in-tech-statistics-the-hard-truths-of-an-uphill-battle.html [accessed 2023-04-01]
41. U.S. Equal Employment Opportunity Commission. Diversity in High Tech. URL: https://www.eeoc.gov/special-report/diversity-high-tech [accessed 2023-02-01]
42. Dimoff JK, Kelloway EK. With a little help from my boss: the impact of workplace mental health training on leader behaviors and employee resource utilization. J Occup Health Psychol 2019 Feb;24(1):4-19. [doi: 10.1037/ocp0000126] [Medline: 29939045]

Abbreviations

CESD-10: 10-item Center for Epidemiologic Studies Depression
CFA: confirmatory factor analysis
GAD-7: 7-item Generalized Anxiety Disorder
IT: information technology
MHL: mental health literacy
MHL-W: Mental Health Literacy Tool for the Workplace
MHSIS: Mental Help Seeking Intention Scale
PSS-10: 10-item Perceived Stress Scale

Edited by A Grover; submitted 27.Feb.2025; peer-reviewed by H Mühlan; revised version received 24.Dec.2025; accepted 09.Jan.2026; published 03.Mar.2026.

Please cite as:
Garcia Colato E, Liu N, Chow A, Sherwood-Laughlin CM, Macy JT
Associations Between IT Job Stressors and Anxiety, Depression, and Stress: Cross-Sectional Study
JMIRx Med 2026;7:e73211
URL: https://xmed.jmir.org/2026/1/e73211
doi:10.2196/73211

Administration Technique of Intranasal Corticosteroid Sprays Among Nepali Pharmacists: Cross-Sectional Study

Amar Prashad Chaudhary[1], PharmD; Suraj Kumar Thakur[2], BPharm; Shiv Kumar Sah[2], BPharm, MPharm

[1]Tribhuvan University Teaching Hospital, Maharajgunj, Kathmandu, Nepal

[2]Institute of Medicine, Maharajgunj Medical Campus, Tribhuvan University, Kathmandu, Nepal

Corresponding Author:
Amar Prashad Chaudhary, PharmD
Tribhuvan University Teaching Hospital, Maharajgunj, Kathmandu, Nepal

Related Articles:

Companion article: http://preprints.jmir.org/preprint/83042

Companion article: https://med.jmirx.org/2026/1/e91439

Companion article: https://med.jmirx.org/2026/1/e91443

Companion article: https://med.jmirx.org/2026/1/e91445

Abstract

Background: Allergic rhinitis is a common condition affecting up to 40% of people worldwide, with a notably high prevalence in South Asia. The primary treatment for moderate to severe allergic rhinitis is intranasal corticosteroid sprays (INCS), the use of which is typically demonstrated to patients by registered pharmacists. However, many patients do not use these sprays correctly.

Objective: This study evaluated the proficiency of pharmacists in demonstrating the correct technique for using INCS and the factors contributing to proper technique.

Methods: In a cross-sectional survey of 365 registered pharmacists in the Kathmandu Valley, Nepal, a trained observer used a standardized 12-step checklist to assess each pharmacist's technique for using INCS. The 12-step checklist was created after studying international guidelines, studies conducted in Nepal, international research articles, and instructional pamphlets. Simple random sampling was done to collect the data from community pharmacies in Kathmandu district. Demographics, education, experience, previous training, and instructional materials use were recorded. A total of 12 marks were awarded for all 12 steps, with one mark given for each step. Proficiency was classified as "adequate" if more than 6 marks were obtained.

Results: Out of 365 pharmacists, 239 (65.5%) were male and 126 (34.5%) were female. Overall, 216 pharmacists (59.2%) were aged 26 years or younger and 235 pharmacists (69.9%) held a diploma in pharmacy. We found that 193 (52.9%) pharmacists demonstrated inadequate technique, while only 172 (47.1%) showed adequate skill overall. However, only 22 pharmacists (6%) demonstrated all 5 critical steps. The likelihood of providing appropriate counseling on the use of INCS was significantly correlated with multiple independent factors. Those with a diploma in pharmacy had a 97% lower likelihood of providing appropriate counseling compared with those with a bachelor's degree in pharmacy and above (P<.001). Pharmacists who perform counseling sessions 1 - 4 times per week had 11-fold greater odds of doing so correctly compared with those who do not (P=.002). Pharmacists who do not use educational leaflets were 96% less likely to provide adequate counseling (P= .005) . Similarly, pharmacists under the age of 26 are 89% less likely than older pharmacists to provide adequate counseling (P=.001). It is interesting to note that men were found to have almost 2.3 times higher odds of providing appropriate counseling than women (P=.02).

Conclusions: More than half of the registered pharmacists in Nepal demonstrated inadequate technique when using INCS. The inadequate patient counseling on INCS use can significantly increase the risk of adverse drug reactions and reduce the efficacy of the therapy. Thus, there is a strong need for educational interventions and policy change for improved proficiency.

(JMIRx Med 2026;7:e83042) doi:10.2196/83042

KEYWORDS

intranasal corticosteroid spray; allergic rhinitis; device use technique; pharmacist; patient counselling, continuing pharmacy education

Introduction

A chronic inflammatory condition of the nasal mucosa, allergic rhinitis (AR) is brought on by immunoglobulin E–mediated responses to allergens breathed in. There are many causes of AR, including pollen, dust mites, cockroach waste, animal dander, fumes and odors, changes in environment, smoke, and certain foods or spices. The most common symptoms of AR are sneezing; stuffy nose; runny nose; itchy nose, throat, eyes, and ears; nosebleeds; clear drainage from the nose; snoring; and breathing through the mouth.

AR affects 10% to 40% of the world's population, and its prevalence is increasing in many countries [1,2]. AR and other allergy disorders are also common in Nepal and the surrounding South Asian nations. A recent school-based study in Nepal, for example, found rhinoconjunctivitis symptoms in 28% of children [3]. AR was responsible for almost 25% of allergy illnesses in Nepal's Gandaki Province [3]. Adolescent AR prevalence in India is estimated at 22%, whereas in adults it was found to be 11% among the general population and 33.3% in asthmatics [4,5]. Similarly, a large-scale study conducted in Europe discovered that up to 20% of the population is impacted by AR [6]. The prevalence of AR in the United States is slightly lower (7.7% in adults and 7.2% in children) [7].

Therefore, the treatment of AR is very important as it impacts daily life activities. The objective of AR treatment is to control the disease. Antihistamines, leukotriene receptor antagonists, azelastine, and intranasal corticosteroid sprays (INCS) are used for treating AR according to the Allergic Rhinitis and its Impact on Asthma guidelines 2019 [8]. Effective pharmacotherapy is crucial for symptomatic control of AR. INCS are the most potent medications for moderate to severe AR and are recommended as first-line therapy [9]. When used correctly, INCS reduce nasal congestion, rhinorrhea, sneezing, and itching by suppressing mucosal inflammation.

The most common adverse drug reactions to INCS include dyspnea, anosmia, ageusia/dysgeusia, epistaxis, and headache [10]. A study conducted at the ear, nose, and throat outpatient clinic at Aberdeen Royal Infirmary found that 15.5% reported epistaxis due to an ipsilateral hand technique [11]. Similarly, a study in Thailand discovered a 3.6 times higher risk of adverse events in patients who did not point the tip of the spray away from the nasal septum [12]. Maintaining a neutral head position and exhaling through the mouth are crucial for proper drug disposition and enhanced efficacy [13]. Therefore, using the correct technique is vital for better efficacy and a reduced risk of side effects. Standard guidelines recommend instructing patients to shake the spray, remove the dust cap, blow the nose, hold the spray bottle while pointing the tip of the nozzle up with the hand, place the index and middle finger on the pusher and the thumb at the bottom of the spray bottle, maintain a neutral head position, insert the tip slightly upward and laterally (away from the septum), close the opposite nostril, inhale gently while actuating the spray, then exhale through the mouth, wipe the nozzle with a tissue or hankerchief, and replace the cap [12,14].

However, a study conducted by Rattanawong et al [12] found that only 4% of patients performed all 12 steps, while only 29% completed all the crucial steps. Similarly, a study by Gurung et al [15] in Nepal revealed that only 7.2% of patients executed all the steps correctly, and 18.2% managed to perform all 5 critical steps accurately (blow the nose, maintain a neutral head position or slightly tilt the head forward, point the tip slightly outward away from the septum, squirt the spray into the nose while breathing in, breathe out through the mouth). A systematic review indicated that approximately 73% of patients did not receive proper advice regarding INCS [16].

Health care professionals, especially pharmacists, are responsible for counseling patients regarding the drugs they dispense. Given this context, it is essential to assess how well Nepali registered pharmacists themselves understand and can demonstrate correct INCS technique. No prior studies have examined this. By identifying gaps in pharmacist knowledge and technique, targeted interventions (eg, curriculum changes or training modules) can be designed to improve AR care. This study therefore evaluated the proficiency of registered pharmacists in Kathmandu Valley in demonstrating INCS administration and analyzed professional factors associated with adequate technique.

Methods

Study Design and Study Period

A cross-sectional observational study was performed from November 1, 2023, to May 28, 2024, through interviews of registered pharmacists. They answered a semistructured questionnaire containing questions about their sociodemographic information, professional details, and INCS counseling steps. STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) principles were adhered to in the study's reporting [17,18].

Study Population and Study Site

The sample was selected from pharmacists registered at the Nepal Pharmacy Council working at community pharmacies registered at the Department of Drug Administration (DDA) in Kathmandu, Nepal. Being Nepal's capital, Kathmandu is a heavily populated city. The respective site had a large number of community pharmacies, about 4000, with many registered pharmacists [19].

Sampling Method

Simple random sampling of the community pharmacies in different wards of Kathmandu district, Nepal, was done using Statistical Package for Social Sciences software (version 26; IBM Corp). The details of all the registered community pharmacies were obtained from the DDA database. No ward-level sampling was performed to avoid geographical clustering.

This cross-sectional study identified potential participants from the registered pharmacists working at community pharmacies. If the community pharmacy had more than one pharmacist, one pharmacist was selected for the study randomly. If the community pharmacy was closed or the pharmacist was not available, a total of three visits were made on different dates; if a pharmacist was still not available, another pharmacy was

selected based on a pregenerated reserved list of random samples. These potential participants were approached and the study's purpose, procedures, and potential risks and benefits were explained to them. The same interviewer interviewed all the participants to overcome interobserver variability in participants' responses.

Sample Size

The survey study was completed using the Raosoft sample size calculator to capture the appropriate sample size [20]. A minimum of 363 samples was required for a 95% confidence interval and a 5% margin of error for the population distribution of 21,000 registered pharmacists at a 40% response distribution [21]. Thus, a total of 365 registered pharmacists participated in this study.

Measures

After the pharmacist's sociodemographic and professional information were obtained through interviews, the 12-step nasal spray application technique as given in Textbox 1 was demonstrated by the participant and examined by the researcher [12,13,22-24]. Each correct step was assigned 1 mark, while incorrect or missed steps were assigned 0 marks. Hence, the maximum score obtained was 12 marks. Five steps in INCS counseling (indicated in Textbox 1) were considered critical based on their impact on patient outcomes and the risk of adverse drug reactions. The median value of the total marks scored was 6.

Textbox 1. Steps for the administration of intranasal corticosteroid sprays.

1. Shake the spray in a vertical plane.
2. Remove the dust cap.
3. Blow the nose (critical).
4. Hold the spray bottle, pointing the tip of the nozzle up with the hand.
5. Place the index and middle finger on the pusher and the thumb at the bottom of the spray bottle.
6. Put the tip of the nozzle into one nostril and close the other side.
7. Maintain a neutral head position or slightly tilt the head forward (critical).
8. Point the tip slightly outward, away from the septum (critical).
9. Squirt the spray into the nose while breathing in (critical).
10. Breathe out through the mouth (critical).
11. Wipe the nozzle with a tissue or handkerchief.
12. Replace the cap.

Determination of the Cutoff Score

To determine the cutoff score, a sensitivity analysis was conducted for alternate cutoffs (ie, >5 and >7). The direction and significance of the main predictors remained stable at >5 and >6, indicating robustness of the findings as shown in Table S1 in Multimedia Appendix 1. The >7 cutoff produced unstable estimates due to small cell sizes. Based on a study conducted by Kc et al [25], expert suggestions, the median value, and sensitivity analysis, more than 6 marks was established as the cutoff score. Therefore, anyone with a score higher than 6 marks was categorized as performing adequately, and anyone with marks equal to or less than 6 was categorized as performing inadequately.

Reliability and Validity

The initial questionnaire was validated by a panel of subject experts, composed of advisors, professors, and teachers, for correctness, clarity, appropriateness, and jargon use. This validation was conducted using face validity approaches. An interrater reliability test was conducted on 15 participants and found a Cronbach α value of 0.758.

Inclusion and Exclusion Criteria

This study only took into account pharmacists aged 18 years and above who were registered with the Nepal Pharmacy Council and employed in community pharmacies. Participants needed to have a Diploma in Pharmacy (DPharm), Bachelor of Pharmacy (BPharm) degree, Doctor of Pharmacy (PharmD) degree, or Master of Pharmacy degree. Participants needed to have a minimum of 1 year of experience. No unregistered pharmacists, pharmacy students, or interns were considered for this study.

Data Collection Procedure

The essential information was then gathered from participants using a semistructured questionnaire administered through an in-person interview. A standardized protocol was followed during interviews. Prior to their enrollment in the study, all participants were informed of its purpose, and their consent was acquired.

Statistical Analysis

Using Microsoft Excel (Microsoft Corp) and Statistical Package for Social Sciences software (version 26; IBM Corp), the gathered data were analyzed. Factors related to the administration technique were evaluated using multivariate binary logistic regression to understand their independent impact. The decision tree analysis was done using Chi-square automatic interaction detector to explore hierarchical relationships and interactions among predictors of INCS

counseling proficiency and to complement the findings of binary logistic regression. When $P<.05$ and the confidence level was 95%, it was deemed statistically significant.

Ethical Considerations

Ethical approval reference number 210 (6-11) E2, 080/081, was provided by the institutional review committee of the Institute of Medicine, Tribhuvan University, before the commencement of the study. Written informed consent was provided by participants before any data were collected from the study site (Multimedia Appendix 2). The identity of participants will not be revealed in any information that will be published or released to third parties. The participants were not compensated for this study.

Results

Participant Characteristics

Pharmacists' professional and demographic traits are listed in Table 1. The study involved 365 registered pharmacists as participants. Of the 365 pharmacists, 216 (59.2%) were ≤26 years old, and 239 were men (65.5%). In addition, 244 (66.8%) were single. Only 110 participants (30.1%) had a BPharm degree or above, whereas 255 (70%) had a DPharm degree. Moreover, 267 participants (73.2%) were early career (1 - 4 y), whereas 98 (26.8%) were mid-career or late career (5 y and above). In all, 194 participants (53.2%) reported counseling patients on intranasal corticosteroids 1 to 4 times per week, but only 30 participants (8.2%) acknowledged any formal training in INCS administration. Additionally, only 75 participants (20.5%) used leaflets to counsel the patients.

Table . Demographic and professional characteristics of pharmacists (N=365).

Variables	Frequency	Percentage
Sex		
Male	239	65.5
Female	126	34.5
Age		
≤26 years	216	59.2
>26 years	149	40.8
Marital status		
Unmarried	244	66.8
Married	121	33.2
Qualification		
DPharm	255	69.9
BPharm and above	110	30.1
Years of experience		
1 - 4 years	267	73.2
5 years and above	98	26.8
Intranasal corticosteroid spray counseling (per week)		
Occasionally	119	32.6
1 - 4 times	194	53.2
More than 4 times	52	14.2
Received training		
Yes	30	8.2
No	335	91.8
Use of information material		
Yes	75	20.5
No	290	79.5

Administration Technique Adherence and Proficiency Level

Among 365 participating pharmacists, adherence to INCS administration steps varied widely, as shown in Table 2. High adherence (>80%) was observed in 4 basic steps: removing the dust cap, replacing the cap, shaking the spray, and holding the bottle upright. In addition, moderate adherence (40% - 80%) was noted for 3 steps: inhaling while spraying, finger positioning, and nozzle insertion. However, low adherence

(<40%) was observed for 5 steps, of which 4 were critical: blowing the nose, pointing the nozzle away from the septum, exhaling through the mouth, proper head positioning, and wiping the nozzle after use.

Table . Performance of each administration step by pharmacists (N=365).

Step	Steps for the administration of intranasal corticosteroid spray	Frequency	Percentage
1	Shake the spray in a vertical plane	309	84.7
2	Remove the dust cap	365	100
3	Blow the nose (critical)	39	10.7
4	Hold the spray bottle, pointing the tip of the nozzle up with the hand	293	80.3
5	Place the index and middle finger on the pusher and the thumb at the bottom of the spray bottle	220	60.3
6	Put the tip of the nozzle in one nostril and close the other side	146	40
7	Maintain a neutral head position or slightly tilt the head forward (critical)	122	33.4
8	Point the tip slightly outward, away from the septum (critical)	36	9.9
9	Squirt the spray into the nose while breathing in (critical)	287	78.6
10	Breathe out through the mouth (critical)	43	11.8
11	Wipe the nozzle with a tissue or handkerchief	123	33.7
12	Replace the cap	359	98.4

The participants' median score across all 12 steps was 6. However, the 5 crucial steps only had a mean score of 1.9 (SD 1.09). Twelve points were awarded for completing all INCS counseling steps, of which 5 points were awarded for the 5 critical steps. Just 22 participants (6%) were able to accurately complete all 5 critical steps. We found that 193 (52.9%) of the registered pharmacists were inadequately knowledgeable on INCS patient counseling. Only 172 participants (47.1%) had adequate knowledge of INCS counseling.

Factors Associated With Proper Administration Technique

Several professional and sociodemographic factors were shown to be substantially correlated with the degree of administration technique proficiency by the multivariate binary logistic regression analysis (Table 3). Years of experience, training, and marital status did not show statistically significant relationships, while sex, age, qualification, frequency of patient counseling weekly, and the utilization of information material were found to be significant predictors.

The likelihood of male pharmacists exhibiting proper technique was about 2 times higher than that of female pharmacists (P=.02). The probability of using an inappropriate INCS counseling technique was 89% lower for individuals who were older than 26 years (P=.001). Proficiency was substantially predicted by having used educational materials. Pharmacists who used educational materials were 96% less likely to perform inadequately (P=.005). Pharmacists with a BPharm degree or higher were also around 97% less likely to counsel inappropriately than those with a DPharm (P<.001). According to this study, individuals who advise patients on INCS 1 - 4 times per week were 11 times more likely to demonstrate proficiency as opposed to those who counsel occasionally (P=.002).

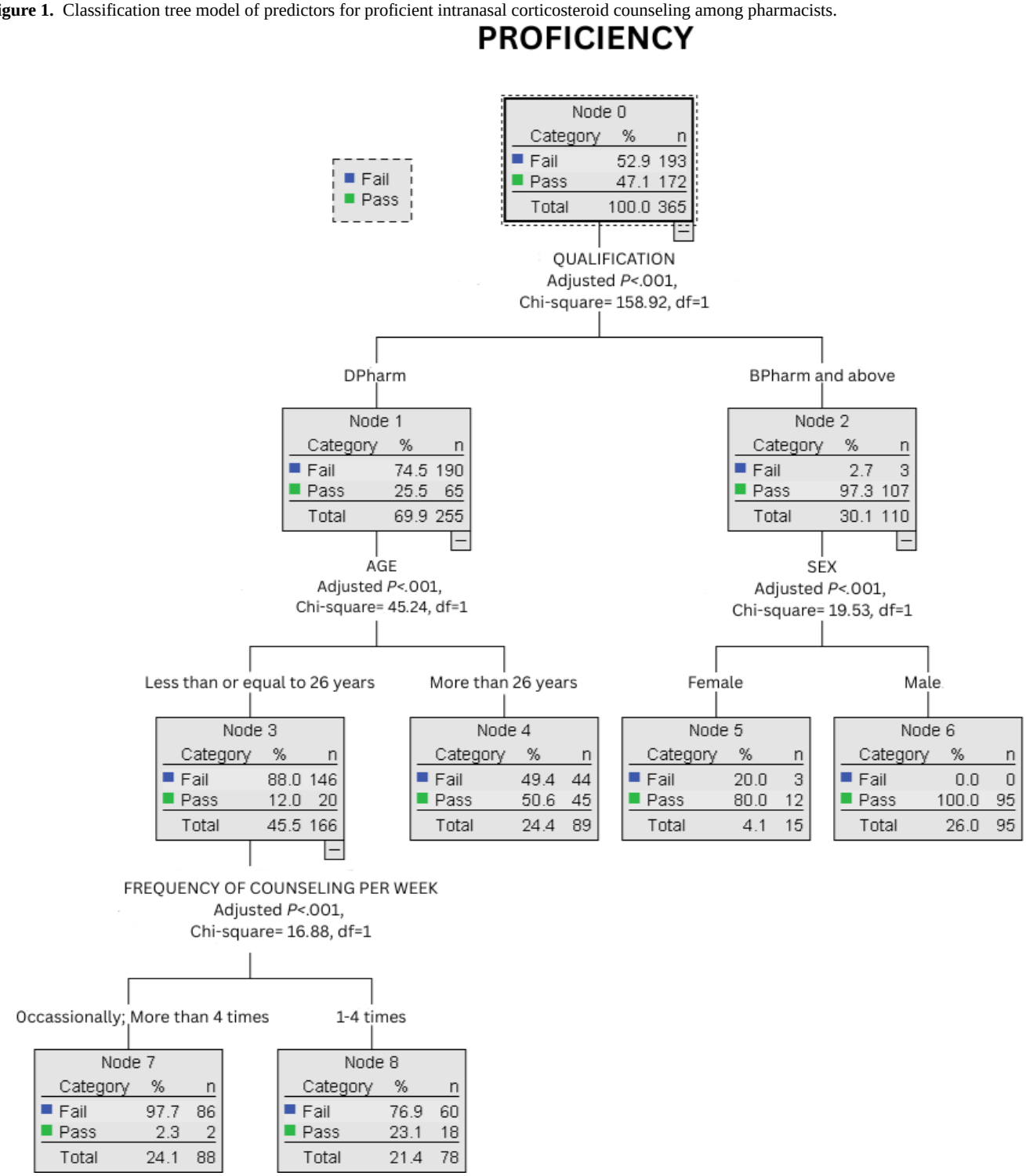
The classification tree (Chi-square automatic interaction detector method), as shown in Figure 1, was developed to identify key predictors of pharmacist proficiency in INCS counseling. The final pruned classification included 5 levels with 9 terminal nodes, achieving an overall classification accuracy of 81.6%. The root node shows the entire study population, and subsequent splits identify variables that best differentiate proficiency levels. Terminal nodes represent final subgroups, displaying the proportion of pharmacists classified as proficient or nonproficient within each subgroup.

Table . Binary logistic regression analysis of proficiency level of administration technique and different sociodemographic and professional details variables.

Variable	Adjusted odds ratio	95% CI	*P*
Sex			
Male	2.30	1.11 - 4.75	.02
Female	Reference		
Age			
≤26 years	0.11	0.03 - 0.41	.001
>26 years	Reference		
Marital status			
Unmarried	2.39	0.71 - 8.06	.16
Married	Reference		
Training			
No	__[a]	—	>.99
Yes	Reference		
Use of educational leaflet			
No	0.04	0.004 - 0.38	.005
Yes	Reference		
Qualification			
DPharm	0.03	0.007 - 0.14	<.001
BPharm and above	Reference		
Years of experience			
1 - 4 years	0.80	0.33 - 1.94	.62
5 years and above	Reference		
Intranasal corticosteroid spray counseling (weekly)			
Occasionally	4.80	0.91 - 25.30	.06
1-4 times	11.21	2.35 - 53.53	.002
>4 times	Reference		

[a]Not applicable.

Figure 1. Classification tree model of predictors for proficient intranasal corticosteroid counseling among pharmacists.



The first and the most significant split was based on the participants' educational qualifications. Only 65 of 255 (pass rate: 25.5%) pharmacists with a DPharm degree had adequate proficiency. However, of 110 pharmacists with a BPharm degree or higher, 107 had adequate proficiency (pass rate: 97.3%). Among the DPharm group, age was another significant predictor. Among those aged less than 26 years, only 20 of 166 participants (12%) had adequate proficiency, whereas among the older peers, 45 of 89 had adequate proficiency. Similarly, among the BPharm group, another major factor was gender. Male pharmacists were found to be 100% proficient in INCS counseling, with all 95 participants demonstrating adequate proficiency, whereas only 12 of 15 female participants had adequate proficiency.

Finally, for younger DPharm degree holders (≤26 y old), the frequency of INCS counseling was another predictor. Those younger participants who counseled occasionally or more than 4 times per week had significantly lower proficiency (2/88 had adequate proficiency) compared to those who counseled 1-4 times per week (18/78 had adequate proficiency).

Discussion

Principal Findings

This study addresses a critical gap in pharmacist competency regarding INCS within resource-constrained health systems, where pharmacists are front-line care providers. This study is among the first in Nepal to assess pharmacists' proficiency with INCS counseling. The survey revealed a significant gap in the participants' understanding of INCS counseling, which helps in understanding its impact on the health outcomes of patients. Approximately 50% of the pharmacists lacked adequate INCS counseling abilities. According to this study, only 6% of pharmacists were able to complete all the essential patient counseling steps that are crucial for appropriate drug administration and to minimize the risk of adverse drug reactions. Classification tree analysis showed that educational degree was the primary predictor of INCS counseling proficiency. Those with BPharm degrees or higher were far more proficient than DPharm degree holders.

The survey's conclusions about the inadequate INCS administration abilities of Nepali registered pharmacists are in line with the findings of patients and medical professionals worldwide [2,14,26]. Only 22 of 365 of pharmacists (6%) performed all recommended steps correctly, which was similar to a study of health care workers in Thailand [14]. However, even in a developed country like the Netherlands, it was found that only about 36% of health care workers were able to complete all the critical steps [26]. These observations suggest that there is a major gap in skill related to INCS counseling across nations, rather than it being a local issue. Due to this inadequate proficiency among pharmacists, there is a high risk of an increase in adverse drug reactions in patients. Therefore, the educational system must be improved to include simulation-based training and mandatory hands-on workshops that allow students and professionals to practice essential steps repeatedly and understand their rationale.

The high proportion of pharmacists demonstrating steps 1, 2, 4, and 12 correctly (>80%) likely reflects common-sense knowledge (shake, remove dust cap, hold the bottle, replace the cap) that is often taught in basic therapy discussions. However, steps like bending the head forward or cleaning the nozzle were rarely done correctly (<40%). This may cause improper drug disposition, irritation in the throat, and increased risk of contamination [27]. Similarly, only about 10% of participants were counseled about pointing the nozzle away from the nasal septum, which reduces the risk of nasal irritation, dryness, and epistaxis, and improves drug absorption from the lateral nasal wall [12,27]. In addition, the steps necessary to remove mucus or debris or obstruction from the nose and reduce throat irritation (ie, blowing the nose before use and exhaling through the mouth) were only performed by about 10% of participants [12]. Patients who are not taught to clean the spray tip may experience clogging or contamination.

These differences align with prior studies indicating that procedural complexity and a lack of continuing pharmacy education (CPE) or training contribute to inconsistent adherence to medical device protocols [28]. This study highlights that even pharmacists, who are trained professionals, often lack full mastery of device use and suggests there is a need to improve the pharmacy curriculum and landscape of CPE in Nepal.

One of the important differences was the pharmacist's qualification. BPharm graduates were about 97% less likely to demonstrate incorrect technique than DPharm graduates. The latter finding reflects the differences in Nepal's educational system. The 3-year DPharm program in Nepal has traditionally emphasized dispensing skills, whereas the BPharm and PharmD curricula include more clinical training.

Shrestha et al [29] found that Nepal's conventional pharmacy education is mostly lecture-based and industry-oriented, with limited practical training in hospitals. Bhuvan et al [30] also documented the challenges in transitioning to PharmD in Nepal, with a focus on patient care and pharmaceutical care. This highlights a need for a gradual change in current policy. Medical devices training should be included in the DPharm degree, and seminars and workshops should involve DPharm students and graduates. Pharmacy regulators in Nepal, such as the Nepal Pharmacy Council or the DDA, may consider upgrading community pharmacists' credentials or introducing minimum competency assessments for patient counseling.

In this study, it was found that pharmacists who used educational leaflets were much more proficient. This is similar to findings of other studies where pharmacist-led interventions with practical demonstrations and the use of leaflets dramatically improved patient technique [25,31]. These educational leaflets significantly reduce the cognitive load of pharmacists and ensure the completeness of all steps. These aids also engage patients through teach-back, reinforce learning beyond completeness, and boost the pharmacist's confidence and professionalism. Therefore, pharmacists should be encouraged to use educational leaflets during counseling sessions on INCS use.

In our study, increasing age (>26 y) was significantly associated with improved INCS counseling proficiency. A study conducted in Korea also found that proficiency in patient counseling regarding topical corticosteroids significantly improved with increasing age [32]. Thus, suggesting increased clinical exposure, more trainings, mature communication skills, and more frequent patient interaction may contribute to better proficiency. In order to succeed in INCS counseling, younger pharmacists must receive sufficient training throughout their time in pharmacy school. They should also attend workshops on medical devices, communication techniques, and patient counseling.

Interestingly, participants counseling on INCS use 1 - 4 times per week have a much higher proficiency (almost 11 times higher) compared with that of participants counseling only occasionally. This relationship likely reflects that a moderate counseling volume provides sufficient repetition to hone skills and confidence, while excessive patient load and task interruptions may reduce time for careful demonstration and feedback [33,34]. Simulation training could help low-counseling pharmacists achieve similar proficiency without relying on clinical exposure.

The analysis of this survey revealed that, among BPharm graduates, males have about 2 times higher odds of proficiency than females regarding INCS counseling. However, the existing literature does not present any conclusive or consistent evidence of sex-based differences in nasal spray or inhalation administration technique among pharmacists. Therefore, the observed difference may reflect contextual, educational, or practice-related factors rather than true gender-based differences.

In our study, 335 of 365 pharmacists (91.8%) lacked specific training. This suggests that continuing professional development for pharmacists in Nepal is sorely needed. According to a recent analysis of continuing professional development in Nepal, CPE is still in its infancy; therefore, working pharmacists are not informed of the latest treatments or best practices [35]. Establishing regular INCS technique workshops or integrating device training into the curriculum could narrow the gap. Given pharmacists' accessibility in rural and urban Nepal [36,37], such measures could rapidly propagate correct practice.

Pharmacists' poor INCS technique skills are concerning but can be resolved. Targeted training in Nepal could help pharmacists improve their skills quickly. Emphasizing AR and device technique in undergraduate pharmacy programs and requiring competency demonstrations during examinations could have a lasting impact. In addition, public health campaigns might encourage patients to ask pharmacists for a demonstration of INCS technique. In the long term, strengthening pharmacy education and integrating pharmacists into asthma/allergy care pathways will benefit Nepal's health care system by improving primary-level management of chronic respiratory diseases.

Limitations

This study was conducted in urban Kathmandu, so findings may not generalize to rural areas, where pharmacies are fewer and mainly operated by trained dispensers. This study has a cross-sectional design, so it cannot prove causality. Potential confounders like workload details were not measured, which may have partly contributed to the large adjusted odds ratio of some predictors. Some of the findings may be extreme due to small subgroups such as pharmacists who had received formal training or those providing frequent INCS counseling. A small sample count can result in unstable estimates and inflate the results. In addition, model overfitting can occur due to the inclusion of multiple interrelated predictors during logistic regression. This study used a small sample size for the reliability test and only used an expert-based face validity test, which may limit the robustness and generalizability of the study. Even with anonymized, behavior-focused questions, self-reported variables like the frequency of counseling and the usage of educational leaflets may be overestimated due to recall and social desirability bias, especially in in-person interviews. This shortcoming is highlighted and the necessity of objective assessment is supported by the observed difference between overall self-reported sufficiency and inadequate performance on critical steps. We recommend a weighted or competency-based scoring model in future studies. Finally, the presence of an interviewer might have influenced the participants' performance (ie, the Hawthorne effect), possibly inflating technique scores. However, due to the low proficiency observed among the participants, any such effect was limited.

Conclusion

This study highlights that more than half of the participants did not have adequate skills to demonstrate proper INCS usage technique. This can lessen its effectiveness in treating AR and increase the likelihood of adverse drug reactions in patients, such as dyspnea, anosmia, ageusia/dysgeusia, epistaxis, and headache. The lack of knowledge is mainly due to poor exposure to this topic in pharmacy school. In addition, training and seminars are limited both during school and after registration as a pharmacist. Resolving this problem should be one of the most important tasks for the Nepal Pharmacy Council and the Health Ministry as AR is very common in Nepal. Upgrading pharmacy curricula, mandating continuing education, and providing standardized counseling materials may empower pharmacists to counsel patients on the correct technique.

Acknowledgments

We would like to express our sincere gratitude to the professors, teachers, and academic staff of Tribhuvan University for their immense support and guidance. We are also thankful to the pharmacists of Tribhuvan University Teaching Hospital and our beloved friends.

Data Availability

The data that support the findings of this study are uploaded here [38].

Authors' Contributions

Conceptualization: APC, SKT
Methodology: APC, SKT
Formal analysis and investigation: APC
Writing – original draft preparation: APC, SKT
Writing – review and editing: APC
Supervision: SKS

Conflicts of Interest

None declared.

Multimedia Appendix 1
Sensitivity analysis.
[DOCX File, 17 KB - xmed_v7i1e83042_app1.docx]

Multimedia Appendix 2
Informed consent form.
[PDF File, 100 KB - xmed_v7i1e83042_app2.pdf]

References

1. García-Almaraz R, Reyes-Noriega N, Del-Río-Navarro BE, et al. Prevalence and risk factors associated with allergic rhinitis in Mexican school children: Global Asthma Network Phase I. World Allergy Organ J 2021 Jan;14(1):100492. [doi: 10.1016/j.waojou.2020.100492] [Medline: 34659624]
2. Akhouri S, House SA, Doerr C. Allergic Rhinitis (Nursing): StatPearls Publishing. URL: https://www.ncbi.nlm.nih.gov/sites/books/NBK568690 [accessed 2023-07-16]
3. Nepali R, Sigdel B, Dubey T, Kc N, Gurung B, Baniya P. Common allergens in patients of allergic rhinitis in Gandaki province of Nepal. JGMC Nepal 2022 Jul 25;15(1):32-36. [doi: 10.3126/jgmcn.v15i1.42141]
4. Moitra S, Mahesh PA, Moitra S. Allergic rhinitis in India. Clin Experimental Allergy 2023 Jul;53(7):765-776. [doi: 10.1111/cea.14295]
5. Sinha B, Vibha, Singla R, Chowdhury R. Allergic rhinitis: a neglected disease - a community based assessment among adults in Delhi. J Postgrad Med 2015;61(3):169-175. [doi: 10.4103/0022-3859.159418] [Medline: 26119436]
6. Bauchau V, Durham SR. Prevalence and rate of diagnosis of allergic rhinitis in Europe. Eur Respir J 2004 Nov;24(5):758-764. [doi: 10.1183/09031936.04.00013904] [Medline: 15516669]
7. Allergic rhinitis: practice essentials, background, pathophysiology. Medscape. URL: https://emedicine.medscape.com/article/134825-overview#a6 [accessed 2025-06-01]
8. Klimek L, Bachert C, Pfaar O, et al. ARIA guideline 2019: treatment of allergic rhinitis in the German health system. Allergo J Int 2019 Nov;28(7):255-276. [doi: 10.1007/s40629-019-00110-9]
9. Neffen H, Wingertzahn MA. Ciclesonide, a hypotonic intranasal corticosteroid. Allergy Asthma Proc 2010;31 Suppl 1:S29-S37. [doi: 10.2500/aap.2010.31.3348] [Medline: 20557684]
10. Ahsanuddin S, Povolotskiy R, Tayyab R, et al. Adverse events associated with intranasal sprays: an analysis of the food and drug administration database and literature review. Ann Otol Rhinol Laryngol 2021 Nov;130(11):1292-1301. [doi: 10.1177/00034894211007222] [Medline: 33813873]
11. Ganesh V, Banigo A, McMurran AEL, Shakeel M, Ram B. Does intranasal steroid spray technique affect side effects and compliance? Results of a patient survey. J Laryngol Otol 2017 Nov;131(11):991-996. [doi: 10.1017/S0022215117002080] [Medline: 29050548]
12. Rattanawong S, Wongwattana P, Kantukiti S. Evaluation of the techniques and steps of intranasal corticosteroid sprays administration. Asia Pac Allergy 2022 Jan;12(1):e7. [doi: 10.5415/apallergy.2022.12.e7] [Medline: 35174058]
13. Benninger MS, Hadley JA, Osguthorpe JD, et al. Techniques of intranasal steroid use. Otolaryngol Head Neck Surg 2004 Jan;130(1):5-24. [doi: 10.1016/S0194-5998(03)02085-0] [Medline: 14726906]
14. Rollema C, van Roon EN, de Vries TW. Inadequate quality of administration of intranasal corticosteroid sprays. J Asthma Allergy 2019;12:91-94. [doi: 10.2147/JAA.S189523] [Medline: 31040706]
15. Gurung U, Khadgi S. Intranasal corticosteroid spray usage in patients with allergic rhinitis: correctness in technique and compliance. J Inst Med Nepal 2024;46(1):1-6. [doi: 10.59779/jiomnepal.1308]
16. Al-Taie A. A systematic review for improper application of nasal spray in allergic rhinitis: a proposed role of community pharmacist for patient education and counseling in practical setting. Asia Pac Allergy 2025 Mar;15(1):29-35. [doi: 10.5415/apallergy.0000000000000173] [Medline: 40051424]
17. STROBE - Strengthening the reporting of observational studies in epidemiology. URL: https://www.strobe-statement.org [accessed 2024-05-30]
18. Elm EV, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP. Strengthening the reporting of observational studies in epidemiology (STROBE) statement: guidelines for reporting observational studies. BMJ 2007 Oct 20;335(7624):806-808. [doi: 10.1136/bmj.39335.541782.AD]
19. Search pharmacy. Department of Drug Administration, Ministry of Health and Population, Government of Nepal. URL: https://dams.dda.gov.np/manLogin/searchPharmacy [accessed 2025-05-04]
20. Sample size calculator. Raosoft Inc. URL: http://www.raosoft.com/samplesize.html [accessed 2024-11-10]
21. Nepal Pharmacy Council. URL: https://nepalpharmacycouncil.org.np/ [accessed 2025-05-03]

22. Rollema C, van Roon EN, van Boven JFM, et al. Pharmacology, particle deposition and drug administration techniques of intranasal corticosteroids for treating allergic rhinitis. Clin Experimental Allergy 2022 Nov;52(11):1247-1263. [doi: 10.1111/cea.14212]
23. How and when to use mometasone nasal spray. NHS. URL: https://www.nhs.uk/medicines/mometasone-nasal-spray/how-and-when-to-use-mometasone-nasal-spray/ [accessed 2025-05-05]
24. Rollema C, van Roon EM, Schilder AG, de Vries TW. Evaluation of instructions in patient information leaflets for the use of intranasal corticosteroid sprays: an observational study. BMJ Open 2019 Jan 15;9(1):e026710. [doi: 10.1136/bmjopen-2018-026710] [Medline: 30647049]
25. Kc B, Khan GM, Shrestha N. Nasal spray use technique among patients attending the out-patient department of a tertiary care hospital, Gandaki Province, Nepal. Integr Pharm Res Pract 2020;9:155-160. [doi: 10.2147/IPRP.S266191] [Medline: 33062617]
26. de Boer M, Rollema C, van Roon E, Vries TD. Observational study of administering intranasal steroid sprays by healthcare workers. BMJ Open 2020 Aug 30;10(8):e037660. [doi: 10.1136/bmjopen-2020-037660] [Medline: 32868363]
27. Intranasal spray technique. National Asthma Council Australia. URL: https://www.nationalasthma.org.au/living-with-asthma/resources/health-professionals/information-paper/intranasal-spray-technique [accessed 2025-05-08]
28. Bosnic-Anticevich SZ, Sinha H, So S, Reddel HK. Metered-dose inhaler technique: the effect of two educational interventions delivered in community pharmacy over time. J Asthma 2010 Apr;47(3):251-256. [doi: 10.3109/02770900903580843] [Medline: 20394511]
29. Shrestha S, Shakya D, Palaian S. Clinical pharmacy education and practice in Nepal: a glimpse into present challenges and potential solutions. Adv Med Educ Pract 2020;11:541-548. [doi: 10.2147/AMEP.S257351] [Medline: 32884392]
30. Bhuvan KC, Subish P, Mohamed Izham MI. PharmD education in Nepal: the challenges ahead. Am J Pharm Educ 2011 Mar 10;75(2):38c. [doi: 10.5688/ajpe75238c] [Medline: 21519429]
31. Chew CC, Lim XJ, Letchumanan P, George D, Rajan P, Chong CP. The effectiveness of pharmacist-led educational model in adult patients with allergic rhinitis: a single-center randomized control trial protocol (AR-PRISE RCT). Trials 2024 Apr 25;25(1):279. [doi: 10.1186/s13063-024-08111-y] [Medline: 38664701]
32. Kang MJ, Park JH, Park S, et al. Community pharmacists' knowledge, perceptions, and practices about topical corticosteroid counseling: a real-world cross-sectional survey and focus group discussions in Korea. PLoS ONE 2020;15(7):e0236797. [doi: 10.1371/journal.pone.0236797] [Medline: 32726366]
33. Shao SC, Chan YY, Lin SJ, et al. Workload of pharmacists and the performance of pharmacy services. PLoS One 2020;15(4):e0231482. [doi: 10.1371/journal.pone.0231482] [Medline: 32315319]
34. Lea VM, Corlett SA, Rodgers RM. Workload and its impact on community pharmacists' job satisfaction and stress: a review of the literature. Int J Pharm Pract 2012 Aug;20(4):259-271. [doi: 10.1111/j.2042-7174.2012.00192.x] [Medline: 22775522]
35. Khatiwada AP, Shrestha S, Sapkota B, et al. Continuing pharmacy education: exploring the status and future prospects in Nepal. Adv Med Educ Pract 2022;13:419-425. [doi: 10.2147/AMEP.S353455] [Medline: 35509353]
36. Lourenço O, Bosnic-Anticevich S, Costa E, et al. Managing allergic rhinitis in the pharmacy: an ARIA guide for implementation in practice. Pharmacy (Basel) 2020 May 16;8(2):85. [doi: 10.3390/pharmacy8020085] [Medline: 32429362]
37. Ikhile I, Anderson C, McGrath S, Bridges S. Is the global pharmacy workforce issue all about numbers? Am J Pharm Educ 2018 Aug;82(6):6818. [doi: 10.5688/ajpe6818] [Medline: 30181678]
38. Study dataset. Zenodo. URL: https://doi.org/10.5281/zenodo.15767865 [accessed 2026-01-27]

Abbreviations

AR: allergic rhinitis
BPharm: Bachelor of Pharmacy
CPD: continuing professional development
CPE: continuing pharmacy education
DDA: Department of Drug Administration
DPharm: Diploma in Pharmacy
INCS: intranasal corticosteroid spray
PharmD: Doctor of Pharmacy
STROBE: Strengthening the Reporting of Observational Studies in Epidemiology

Edited by A Schwartz; submitted 26.Aug.2025; peer-reviewed by RP Shankar, S Chi Lik Au; revised version received 07.Jan.2026; accepted 12.Jan.2026; published 29.Jan.2026.

Please cite as:
Chaudhary AP, Thakur SK, Sah SK
Administration Technique of Intranasal Corticosteroid Sprays Among Nepali Pharmacists: Cross-Sectional Study
JMIRx Med 2026;7:e83042
URL: https://xmed.jmir.org/2026/1/e83042
doi:10.2196/83042

Automating Individualized Notification of Drug Recalls to Patients: Complex Challenges and Qualitative Evaluation

Meghana Gadgil[1], MD, MPH; Rose Pavlakos[2], PharmD; Simona Carini[1], MA; Brian Turner[3], MBA; Ileana Elder[4], PhD; William Hess[4], BS; Lisa Houle[5], BA; Lavonia Huff[4]; Elaine Johanson[4], BS; Carole Ramos-Izquierdo[4], MS, MPM; Daphne Liang[4], PharmD; Pamela Ogonowski[4], MLS (ASCP); Joshua Phipps[6]; Tyler Peryea[4], BA; Ida Sim[1,3], MD, PhD

[1]Division of General Internal Medicine, University of California, San Francisco, Box 0320, San Francisco, CA, United States

[2]Division of Cardiology, University of California, San Francisco, San Francisco, CA, United States

[3]Clinical and Translational Science Institute, University of California, San Francisco, San Francisco, CA, United States

[4]Food and Drug Administration, Silver Spring, MD, United States

[5]Tuvli, LLC, Herndon, VA, United States

[6]Conceptant, Inc, Falls Church, VA, United States

Corresponding Author:
Simona Carini, MA
Division of General Internal Medicine, University of California, San Francisco, Box 0320, San Francisco, CA, United States

Related Articles:

Companion article: https://www.medrxiv.org/content/10.1101/2024.09.18.24312141v1

Companion article: https://med.jmirx.org/2026/1/e82612

Companion article: https://med.jmirx.org/2026/1/e82613

Companion article: https://med.jmirx.org/2026/1/e82609

Abstract

Background: Consumer-level drug recalls usually require action by individual patients. The Food and Drug Administration (FDA) has public-facing outlets to inform the public about drug safety information, including all recalls, but individual consumers may not be aware of them. And there is no system in place to notify individual prescribers which of their patients are affected by a specific recall.

Objective: We aimed to leverage the FDA's Healthy Citizen prototype web-based software platform, which provides users with information about recalls, to automatically notify patients of relevant recalls.

Methods: We developed and evaluated an electronic notification system in the primary care and cardiology practices at a large, urban, academic medical center. The health care portal scanned the FDA Healthy Citizen application programming interface nightly to detect new recalls, identified patients who had those medications in their electronic health record (EHR) medication list, and sent them a message through the EHR patient portal with a link to a customized FDA information display. Using structured interviews, we assessed qualitative feedback on the system and portal messaging from a convenience sample of 9 patients.

Results: The system was technically functional, but it was not possible to trace a medication prescription from the EHR to specific lot numbers dispensed to that patient by a community pharmacy. The qualitative feedback obtained from patients showed convergence of topics.

Conclusions: Lack of an accurate electronic audit trail from prescription to dispensed medication precludes clinical deployment of automated drug recall notification.

(JMIRx Med 2026;7:e68345) doi:10.2196/68345

KEYWORDS

notification system; drug recalls; patient safety; medication; electronic health records; prescriptions; decision support

Introduction

Background

In the United States, the Food and Drug Administration (FDA) is responsible for assuring the safety and efficacy of marketed drugs. When a safety concern arises on a marketed drug, communicating this information to patients is essential, and timely clinical action by prescribers is often required. Yet, patients and prescribers often lack relevant, timely information, leaving patients and health systems unable to efficiently manage drug recalls and their impacts. Recognizing this problem, the FDA developed prototype technology for patients and health systems to automatically be notified of drug recalls through their health care portals as part of the FDA's Healthy Citizen prototype platform that seeks to allow "citizens and those who care for them, research organizations, and FDA to communicate and collaborate in a single, seamless environment connected through the healthcare portal and leveraging the trusted relationships between providers and patients to improve public health outcomes" [1].

Drug Recall Process

Firms, including manufacturers and own-label distributors, can initiate a recall, either on their own or in response to an FDA recommendation, request, or order. Common reasons for recalls include contamination, mislabeling, adverse reactions, defective products, and incorrect potency [2,3]. The FDA works with firms as they develop their recall strategy, which is dependent on a variety of factors, including, but not limited to, the product's degree of hazard, the ease of identifying the product, and the extent of distribution. Depth of recall is one component of this strategy: consumer-level recalls should be extended to consumers and patients; retail-level recalls affect community pharmacies and health care providers; and wholesale-level recalls affect manufacturers and distributors.

For consumer-level recalls, which were the focus of this project, consumers may learn of a recall through FDA.gov [4], news media, or notification from the recalling firm or pharmacy. (Most pharmacies have protocols in place to handle recalls, which may include outreach to customers.) Consumer notifications often recommend that patients consult their health care provider about the best course of action. However, recalls often affect only certain lots of pills, and prescribers have no way of knowing the lot number of the medication dispensed to the patient and therefore whether the patient is affected. The patient often cannot identify the lot number, either, as most dispensing pharmacies are not required to document the lot number on pill bottle labels (see Principal Findings section for details). Thus, if patients contact their health care providers about a recall, the only action providers can take is to redirect patients to their pharmacy. The pharmacy then either replaces the pills with those from an unaffected lot or, if no substitute is available, notifies the prescribing clinician to issue a new prescription for a different medication, dosage, or formulation.

This partnership with the FDA aimed to address the inefficiencies in recall notification by demonstrating timely, fully automated, and individualized communication of drug recalls and recommended actions to patients.

Methods

Study Setting and Participants

The University of California, San Francisco (UCSF), an academic medical center, partnered with the FDA [5] to demonstrate use of Healthy Citizen tools to automate individualized drug recall notifications to outpatient primary care and cardiology patients.

We developed an electronic notification system and conducted this study in the Division of General Internal Medicine (DGIM) and Division of Cardiology at UCSF, a large, urban, academic medical center in San Francisco, California. The DGIM primary care clinic serves 25,000 patients with approximately 70,000 visits yearly. The cardiology clinics serve over 12,000 patients with approximately 30,000 visits yearly. The clinics use the Epic electronic health record (EHR) with the MyChart patient portal. At the time of the study, approximately 45% of patients had actively used MyChart at least once.

We created and tested the notification system within Epic's ACE6 development environment, with the intent to migrate it to production after successful testing. The project team comprised clinicians and programmers from the medical center, FDA leaders from the Office of Health Informatics and other sections, and developers of the Healthy Citizen platform. For testing purposes, fictitious patients were created in the Epic ACE6 environment with medication lists that contained prescriptions matching fictitious medication recalls issued by the FDA.

The prototype system was shown to a convenience sample of 9 patients via remote videoconferencing to obtain initial formative feedback.

Ethical Considerations

Ethical approval for this program evaluation was obtained from the UCSF Institutional Review Board (19 - 27668). Informed consent was obtained from each participant via electronic signature before the interview. Interviews were conducted by 2 investigators (MG and RP). Transcripts were analyzed for common themes by the same, with additional verification by 2 other investigators (SC and IS). Transcripts were stored in a secure location behind an institutional firewall. No identifying data were shared or presented beyond summary statistics (number and gender of patients). Upon completion of the interview, each patient was paid US $25 for their time.

Results

Program Description

The notification system comprised two major technical parts. The first part, within the medical center's firewalls, checked for new consumer-level drug recalls and notified affected patients via MyChart (see the EHR Build section below). The second part was the FDA's Healthy Citizen prototype platform, which provided an application programming interface (API) for external systems to request the latest drug recall information and mechanisms to launch a widget displaying details about a specific recall (see the Healthy Citizen Build section below).

The widget was a SMART-on-FHIR software module that could be embedded into and accessed within an EHR without the need for any additional sign-in.

EHR Build

The EHR build had three major parts: (1) checking for new drug recalls; (2) matching recalls to the patient medication lists; and (3) preparing and sending personalized MyChart notifications to patients. Each part proved extremely challenging to build for technical and data availability reasons.

First, the system issued a nightly call to the Healthy Citizen API to retrieve the National Drug Codes (NDCs) of newly recalled drugs. The next step, matching recalls to a patient's EHR medication list, can result in false negatives and false positives. False negatives can occur if a patient's prescription is missing from the medication list [6], or if the algorithm fails to detect a true match. False positives can arise from two inaccuracies. Crucially, EHR medication lists contain the *prescribed* drug, not the *dispensed* drug. To identify a prescribed drug, Epic uses RxNorm codes that do not include the manufacturer name. To identify recalled drugs and their manufacturers, the FDA uses NDCs, which are unique, 3-segment numbers that identify a drug's labeler (ie, manufacturer or distributor), product, and trade package size [7]. Thus, for example, the NDC from a lisinopril recall from a specific manufacturer will match the RxNorm code for all lisinopril prescriptions of the same strength, regardless of manufacturer. This will erroneously identify patients who were prescribed lisinopril but were not dispensed pills from the affected manufacturer. Secondly, recalls often involve only specific lots, information that is unavailable in the EHR, thereby contributing to false positives, as discussed above.

The third part of the EHR build was to send a MyChart notification to patients once a match was made, alerting them that they may be taking a recalled medication (Figure 1).

The Medication Recalls link led to the FDA Healthy Citizen's display widget, launched as a new window within MyChart showing details of the matched recall, including affected manufacturers (Figure 2). Because the matching algorithm could not restrict matches to affected manufacturers, the MyChart message asked patients to compare the manufacturer name on their pill bottle's label to the manufacturer or manufacturers listed in the FDA informational display and to call their pharmacy if it matched. The patient advisory council of the primary care clinic provided input on the wording and endorsed the importance of the project aims.

Figure 1. MyChart notification of potentially relevant recall.

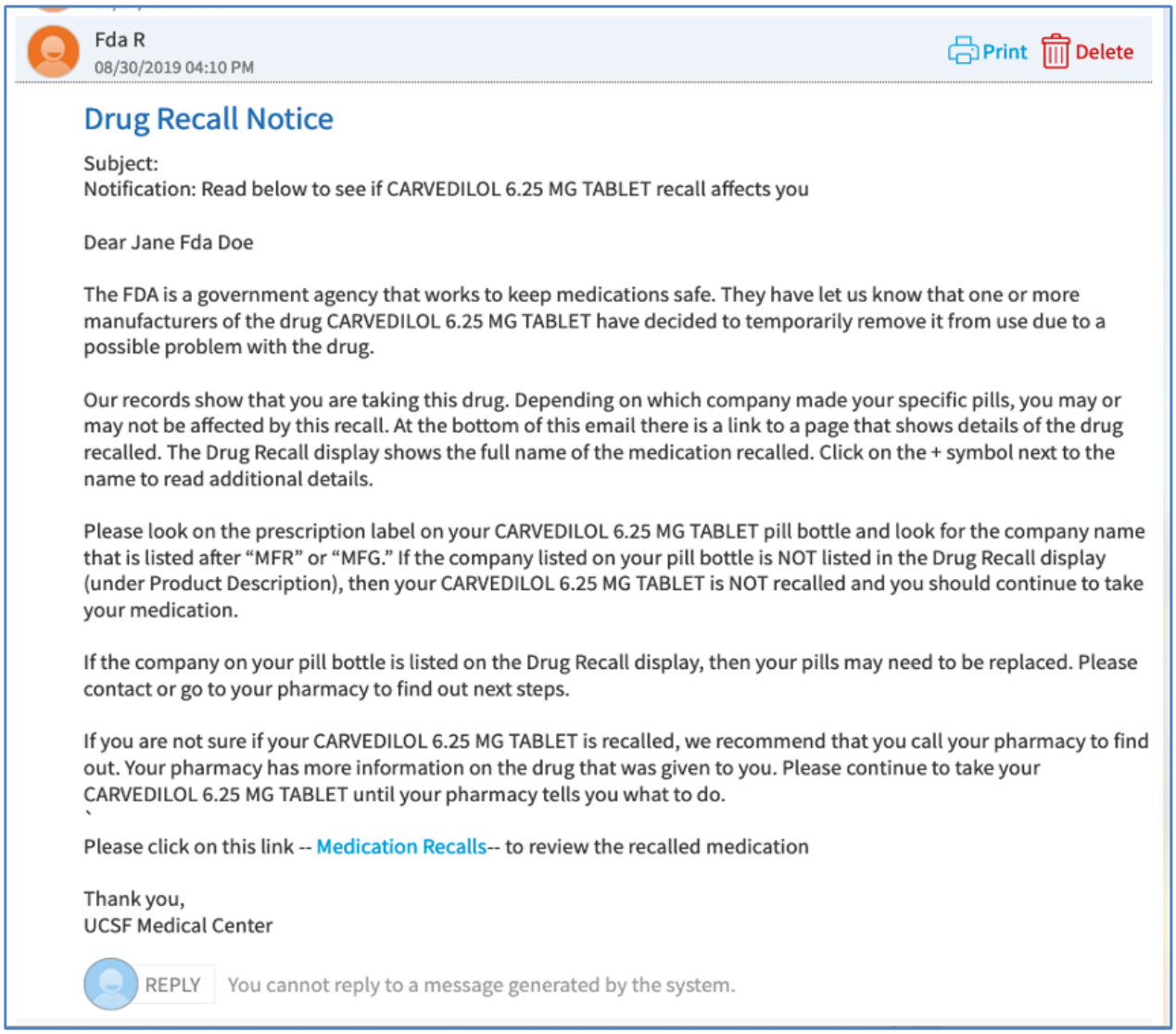
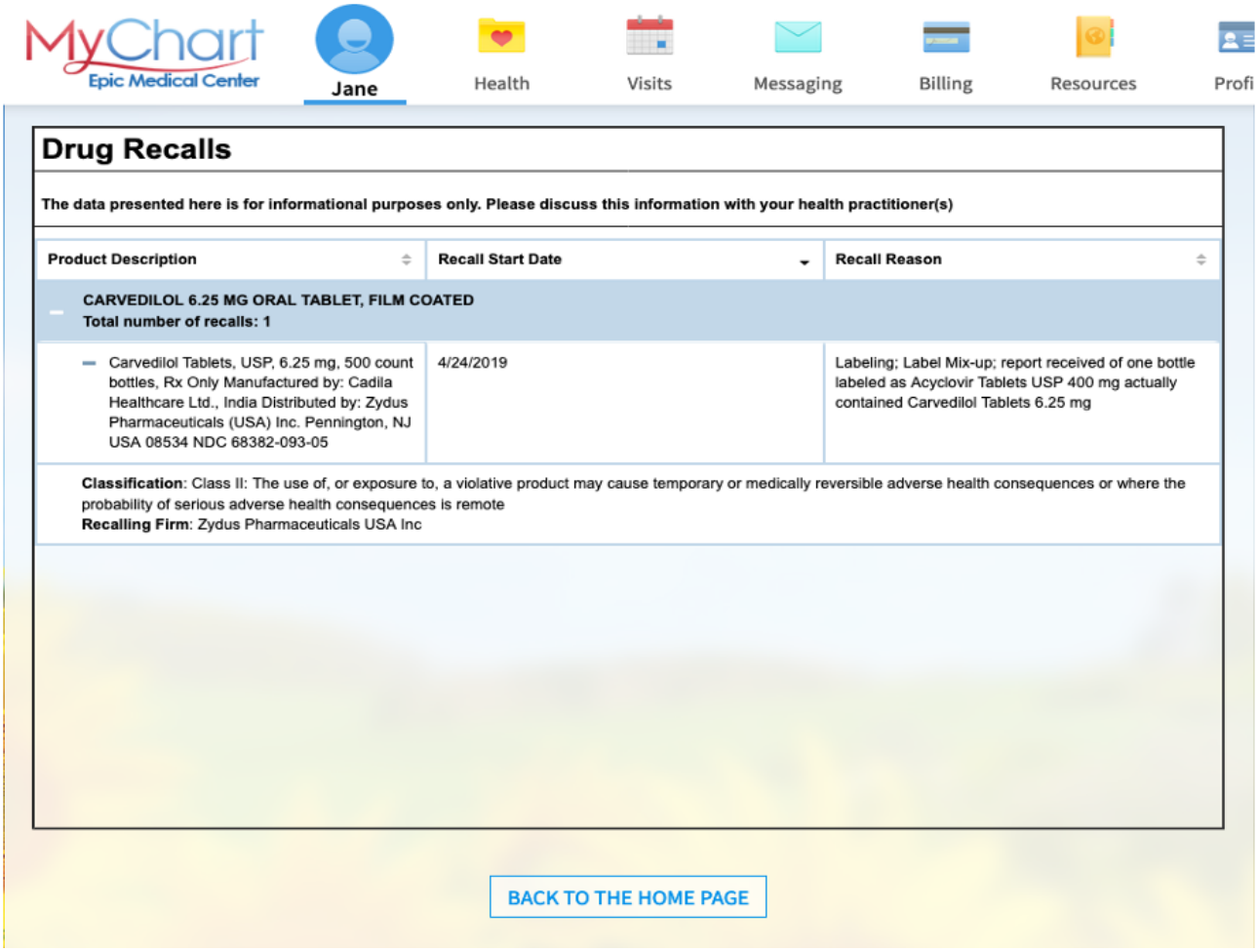


Figure 2. Food and Drug Administration Healthy Citizen information display widget showing official information about a drug recall.



Healthy Citizen Build

Substantial technical work was performed on the Healthy Citizen platform to satisfy the use case needs. For example, technical and internal FDA administrative changes were required to make the depth of recall (ie, retail or consumer level) searchable and to distinguish between new and ongoing recalls. The SMART-on-FHIR widget needed to be available on Epic's App Orchard, and modifications were required for the widget to be called by and launched within Epic. The contents of the widget display were modified to exclude information not relevant to patients, such as the status of the recall (eg, whether it was ongoing or completed), or to move it from the main display to the Additional Details section. The text immediately below the title could not be modified.

Program Evaluation

The system was able to automatically detect a new fictitious medication recall using the Healthy Citizen API, compare and detect matches to each (fictitious) patient's list of prescribed medications, send a MyChart message to affected (fictitious) patients, and launch a display for the correct recall or recalls. The system responded correctly to test patients with zero to multiple affected medications.

Established patients at the primary care clinic who were members of the patient advisory council, used MyChart, and were prescribed at least one medication received a recruitment letter. Patients at the cardiology clinics who were scheduled to see the pharmacist during a random week, who actively used MyChart (or their family members who used MyChart on their behalf), and who were taking at least one prescription medication were deemed eligible for the study and sent a recruitment letter. Interested patients contacted the study team to participate.

We obtained qualitative feedback by interviewing a convenience sample of 9 patients (5 female, 4 male). Two of the 9 participants had personal experience with recalls. Participants were interviewed individually using Zoom (Zoom Communications, Inc). During the session, they were presented with a scenario for fictitious patient Jane Doe, who was prescribed carvedilol (6.25 mg). Using structured interviewing techniques, we evaluated participants' understanding of the MyChart message, widget, and example pill bottle label. Throughout the process we asked for descriptive feedback. Recordings of the interviews were transcribed and separately analyzed by 2 investigators (MG and RP) for common themes [8], with additional verification by SC and IS.

All 9 participants understood the purpose of the MyChart notification message but thought it was too wordy. All 9 were able to identify the medication manufacturer on the example pill bottle label. Only 2 would have clicked on their own on the link at the bottom of the MyChart message to launch the widget; the other 7 needed the interviewer's prompting and guidance

to do so. As advised by the MyChart message, all 9 users would have contacted their pharmacist, but 5 of the 9 would also have contacted their doctor's office, as advised by the widget. Given the choice, all 9 would have liked to receive MyChart notification of potential drug recalls.

Major thematic findings included the following: (1) Patients appreciated being notified of recalls by their clinic, even though their actual medication may not have been affected by the recall, because they trusted the clinic, and the notification showed that the clinic was aware of patient medication issues. (2) Patients saw communicating through the MyChart patient portal as a trusted, efficient, and reliable notification method. Mailed letters can be ignored, and several users said they did not answer phone calls from unknown phone numbers (eg, their pharmacy). (3) Patients suggested that the widget content should be displayed directly in the MyChart message rather than in a new window. (4) Patients felt that the widget itself should be redesigned to more directly meet patient information needs (much of the widget content was either confusing or irrelevant to patients, eg, recall start date, manufacturer address), that the recall reason was appreciated but unnecessary, and that the widget should not ask patients to discuss the information with their health care provider. (5) Patients wanted to discuss the recall with their clinicians to "close the loop."

The project team concluded that operational deployment of this system may lead to unnecessary and unacceptable patient anxiety generated by false positive notifications. In addition, because patient feedback suggested that patients would contact their clinicians regardless of the advice to contact their pharmacy, the system was likely to increase staff burden for responding to patient inquiries. While the project implementation provided important lessons, it did not provide a solid enough business case to justify expanding the pilot, which would have required institutional support. We therefore decided not to proceed with implementation of the FDA drug recall notification system into clinical care.

Discussion

Principal Findings

Drug recalls are an ongoing challenge in the United States [3] and other countries [9]. According to an analysis of FDA recall data, between 2012 and 2023 there were on average 330 recalls per year [3]. When, in 2018, several angiotensin II receptor blockers (prescribed to treat hypertension, heart failure, and chronic kidney disease) were recalled for carcinogenic impurities, the availability of treatments in the same or similar drug class facilitated patients' transition to alternatives [10]. Sustained media attention highlighted communication needs and challenges among the parties impacted.

Patients and clinicians need an accurate system for identifying which patients are affected by which drug recalls and acting on them in a timely and appropriate manner to prevent patient harm and erosion of trust in prescribers and the health care system.

This project demonstrated the technical and clinical feasibility of using the FDA's Healthy Citizen drug recall tools to automatically alert patients, via Epic's patient portal MyChart, to relevant drug recalls. While our project was technically successful, it revealed substantial challenges to responding to drug recalls. Chiefly, while patients want and expect their prescriber to be aware of, and involved in, responding to a drug recall, prescribers have no easy access to the manufacturer and lot number of the actual medication dispensed to their patients. Without these details, health systems cannot accurately target patients and false positive notifications are inevitable. A partial technical solution could be to access Surescripts records, which include the NDC for dispensed drugs as reported to Surescripts via claims data. However, only 70% of patients at UCSF Medical Center use a Surescripts-participating pharmacy, and Surescripts records do not include dispensed lot numbers, such that false positive recall notifications would still be an issue.

Our project showed that a strong case can be made for requiring each pill bottle to include on its label the lot number and NDC of the pills (which links to the manufacturer, labeler, or distributor), so that patients could definitively determine if a recall affected them. Current federal regulation allows such information to appear on an internal leaflet or a label on the outer carton or wrapper of manufactured medications [11], which many patients discard even if the pharmacist includes them with the dispensed medication. As of August 2025, our review of state regulations identified jurisdictions with explicit requirements. Only four state boards of pharmacy (Colorado, Delaware, Oklahoma, and Wyoming), plus the US territory of Puerto Rico, require the lot number to appear on the dispensed medication label [12-16]. In addition, only three state boards of pharmacy (Pennsylvania, New Hampshire, and Ohio) have regulations about the NDC appearing on the dispensed medication label [17-19]. The Pennsylvania State Board of Medicine requires the NDC to appear on the dispensed medication label if the prescriber specifies that the drug name *not* appear on the label [17]. The state boards of pharmacy of New Hampshire and Ohio allow the use of the NDC as an abbreviation for the manufacturer or distributor name, though they do not require it on every dispensed medication label [18,19]. The FDA does not have the legal authority to regulate the practice of pharmacy in any state and therefore cannot require that the lot number and NDC (or anything else, including the name of the drug) be placed on each prescription that a pharmacist dispenses to a patient. The manufacturer and lot number of dispensed medications should routinely be available electronically to prescribing clinicians via standard APIs so that health systems can meet patient expectations that they are trusted guides in properly responding to drug recalls. Policy and data infrastructure changes are required at the regulatory, health IT, and consumer pharmacy levels before automated recall notification can be widely deployed.

Conclusions

The need of patients and clinicians to identify applicable drug recalls and appropriately act on them is currently unmet. Through our project we learned several lessons, which in some cases can be generalized beyond its scope: (1) Patients appreciated receiving a notification showing that the clinic was aware of the patient's medication issues. (2) The MyChart patient portal was seen as a trusted and reliable notification method. (3) Patients preferred the notification content to be

displayed directly in the MyChart message rather than in a new window. (4) Patients considered that the content of the notification should directly address patient information needs, avoiding content that is not strictly necessary. (5) Prescriptions being a sensitive topic, patients wished to discuss the recall with their clinicians, even when directed to contact the dispensing pharmacy.

Our project showed that access to the manufacturer and lot number of the drug dispensed via standard APIs is a requirement for the development and deployment of technical solutions that implement accurate automated recall notifications to patients. While a change at the federal level would be ideal, advocating for individual state boards of pharmacy to require the NDC and lot number to appear on the dispensed medication label may provide needed interim progress for allowing development and deployment of solutions supporting patients' needs.

Acknowledgments

This project was made possible by funding from the US Food and Drug Administration (FDA; grant U01FD005978), which supports the UCSF-Stanford Center of Excellence in Regulatory Sciences and Innovation. The manuscript's contents are solely the responsibility of the authors and do not necessarily represent the official views of the US Department of Health and Human Services (HHS) or the FDA. Any policy recommendations in this manuscript are offered for discussion and do not represent HHS or FDA policy or commitments. This project was funded with US $235,000 (representing 100%) by the FDA/HHS. This work was presented at the Innovations in Regulatory Science Summit in January 2020 and at APhA2020 in March 2020.

Conflicts of Interest

None declared.

References

1. FDA/Healthy-Citizen-Code. Food and Drug Administration. 2020. URL: https://github.com/FDA/Healthy-Citizen-Code [accessed 2020-10-20]
2. Hall K, Stewart T, Chang J, Freeman MK. Characteristics of FDA drug recalls: a 30-month analysis. Am J Health Syst Pharm 2016 Feb 15;73(4):235-240. [doi: 10.2146/ajhp150277] [Medline: 26843501]
3. Ghijs S, Wynendaele E, De Spiegeleer B. The continuing challenge of drug recalls: insights from a ten-year FDA data analysis. J Pharm Biomed Anal 2024 Oct 15;249:116349. [doi: 10.1016/j.jpba.2024.116349] [Medline: 39029352]
4. Recalls, market withdrawals, & safety alerts. US Food and Drug Administration. 2020. URL: https://www.fda.gov/safety/recalls-market-withdrawals-safety-alerts [accessed 2020-10-20]
5. UCSF-Stanford Center of Excellence in Regulatory Science And Innovation (CERSI). University of California San Francisco. URL: https://pharm.ucsf.edu/cersi [accessed 2020-10-20]
6. Tan Y, Elliott RA, Richardson B, Tanner FE, Dorevitch MI. An audit of the accuracy of medication information in electronic medical discharge summaries linked to an electronic prescribing system. Health Inf Manag 2018 Sep;47(3):125-131. [doi: 10.1177/1833358318765192] [Medline: 29587532]
7. National Drug Code Directory. US Food and Drug Administration. 2023 Mar 23. URL: https://www.fda.gov/drugs/drug-approvals-and-databases/national-drug-code-directory [accessed 2025-10-02]
8. Im D, Pyo J, Lee H, Jung H, Ock M. Qualitative research in healthcare: data analysis. J Prev Med Public Health 2023 Mar;56(2):100-110. [doi: 10.3961/jpmph.22.471] [Medline: 37055353]
9. Annema PA, Derijks HJ, Bouvy ML, van Marum RJ. Impact of drug recalls on patients in the Netherlands: a 5-year retrospective data analysis. Clin Pharmacol Ther 2024 Jun;115(6):1365-1371. [doi: 10.1002/cpt.3220] [Medline: 38390768]
10. Callaway Kim K, Roberts ET, Donohue JM, et al. Changes in blood pressure, medication adherence, and cardiovascular-related health care use associated with the 2018 angiotensin receptor blocker recalls and drug shortages among patients with hypertension. J Manag Care Spec Pharm 2025 May;31(5):461-471. [doi: 10.18553/jmcp.2025.31.5.461] [Medline: 40298307]
11. 21 CFR § 201.10 Drugs; statement of ingredients. Code of Federal Regulations. 2019. URL: https://www.ecfr.gov/current/title-21/chapter-I/subchapter-C/part-201/subpart-A/section-201.10 [accessed 2025-07-03]
12. Colorado State Board of Pharmacy Rules and Regulations – 3 CCR 719-1. Colorado Secretary of State. URL: https://www.coloradosos.gov/CCR/GenerateRulePdf.do?ruleVersionId=11936&fileName=3%20CCR%20719-1 [accessed 2025-10-02]
13. Delaware Board of Pharmacy Regulations – 24 del. admin. code § 2500-6.3. Delaware Department of State, Division of Professional Regulation. URL: https://regulations.delaware.gov/AdminCode/title24/2500.shtml [accessed 2025-11-07]
14. Oklahoma Administrative Code, Title 535 Oklahoma State Board of Pharmacy – 535:15-18-4. Oklahoma Secretary of State. URL: https://oklahomarules.blob.core.windows.net/titlepdf/Title_535.pdf [accessed 2025-11-07]
15. Wyoming Pharmacy Act Rules – Wyoming Board of Pharmacy, chapter 2, section 7. Wyoming State Board of Pharmacy. URL: https://rules.wyo.gov/DownloadFile.aspx?source_id=21691&source_type_id=81&doc_type_id=110&include_meta_data=Y&file_type=pdf&filename=21691.pdf&token=062231007102232034219249120076161206179048169170 [accessed 2025-11-07]

16. Laws of Puerto Rico – 20 L of PR § 410a. Justia, US Law, US Codes and Statutes. URL: https://law.justia.com/codes/puerto-rico/title-twenty/chapter-20/subchapter-v/410a/ [accessed 2025-11-07]
17. Pennsylvania Code, State Board of Pharmacy – 49 Pa. Code § 27.18(d). Commonwealth of Pennsylvania. URL: https://www.pacodeandbulletin.gov/Display/pacode?file=/secure/pacode/data/049/chapter27/s27.18.html&d=reduce [accessed 2025-11-07]
18. New Hampshire Code – NH Admin. Rules Ph 703.05(e). New Hampshire Pharmacy Laws & Rules. URL: https://www.oplc.nh.gov/sites/g/files/ehbemt441/files/inline-documents/sonh/nh-phcy-law-rule-book-10-29-19.pdf [accessed 2025-11-07]
19. Ohio Administrative Code – OAC 4729:5-5-06(A)(8). Ohio Laws and Administrative Rules. URL: https://codes.ohio.gov/ohio-administrative-code/rule-4729:5-5-06 [accessed 2025-11-07]

Abbreviations

API: application programming interface
DGIM: Division of General Internal Medicine
EHR: electronic health record
FDA: Food and Drug Administration
NDC: National Drug Code
UCSF: University of California, San Francisco

Edited by CN Hang; submitted 04.Nov.2024; peer-reviewed by A Russ, RC Marshall; revised version received 15.Aug.2025; accepted 18.Aug.2025; published 13.Jan.2026.

Please cite as:
Gadgil M, Pavlakos R, Carini S, Turner B, Elder I, Hess W, Houle L, Huff L, Johanson E, Ramos-Izquierdo C, Liang D, Ogonowski P, Phipps J, Peryea T, Sim I
Automating Individualized Notification of Drug Recalls to Patients: Complex Challenges and Qualitative Evaluation
JMIRx Med 2026;7:e68345
URL: https://xmed.jmir.org/2026/1/e68345
doi:10.2196/68345

Publisher:
JMIR Publications
130 Queens Quay East.
Toronto, ON, M5A 3Y5
Phone: (+1) 416-583-2040
Email: support@jmir.org