Perspective AI Safety, Ethics and Regulation Open access

Demo article. This is a design sample, not real research. The authors, institutions, data and results are fictional, and the dates only illustrate the layout.

Deployed clinical AI needs a denominator: a minimum dataset for monitoring after deployment

Naomi Bergstrand-Ali[1*], Rafael Iturbe[2] and Chidinma Okeke-Lawson[3]

[1] Demo Institute for Health Policy, Washington, DC, USA

[2] Program in Health Law, Demo University School of Law, Washington, DC, USA

[3] Department of Clinical Informatics, Demo University Hospital, Baltimore, MD, USA

*** Corresponding author:** Naomi Bergstrand-Ali, email withheld for this demo.

Received	Revised	Accepted	Published
July 20, 2026	September 7, 2026	September 28, 2026	October 20, 2026

Volume	DOI	License	Peer review
1 (2026), article 6	Not yet assigned	CC BY 4.0, copyright the authors	Not peer reviewed (demo)

Journal	Publisher
Augmented Medicine	Eldenhall Sciences Inc., Harrisonville, MO, United States

Demo dates: the journals opened on November 2, 2026. These dates only show the layout of an article's history.

Cite this article: Bergstrand-Ali N, Iturbe R, Okeke-Lawson C. Deployed clinical AI needs a denominator: a minimum dataset for monitoring after deployment. Augment. Med. 2026;1:6. Available from: https://eldenhall.org/articles/post-deployment-monitoring-clinical-ai-minimum-dataset/ [Demo article]

Abstract

Clinical AI systems are approved and bought on the strength of accuracy figures from development data, and then deployed into settings where almost nothing about their use is measured. Hospitals rarely know what share of eligible patients a model scored, how many alerts it raised, how often clinicians acted on them, or whether its predictions stayed accurate as patients, practice and data changed. We argue that every deployed clinical AI system should report a small set of measures with explicit denominators, covering exposure, output, use, performance, equity, safety and change, and we propose a minimum dataset. We discuss who should collect it, how outcomes contaminated by the model's own effect should be handled, and why the burden is smaller than the cost of not knowing.

DEMO

Keywords: post-market surveillance; clinical decision support systems; dataset shift; patient safety; health policy; quality improvement

1. The missing denominator

A hospital that buys a medical device expects to know how often it is used and how often it fails. For clinical AI, that expectation is rarely met. The number of AI-enabled medical devices cleared by the US Food and Drug Administration has grown quickly [1,2], but an analysis of these clearances found that most devices had been evaluated only on retrospective data, and many without data from more than one site [3]. After a system goes live, the information available to the hospital is usually still the accuracy reported by its developer, measured on someone else's patients, years earlier.

What is missing is the denominator. A sepsis alert that fires 40 times a day tells a hospital nothing on its own. It matters how many patients were eligible and how many were scored, how many of the alerts were for patients who went on to develop sepsis, how many were acted on, and whether these proportions differed between wards and patient groups. The external validation that showed poor performance of a widely deployed proprietary sepsis model was possible because one health system assembled exactly these denominators for its own patients [4,5]. Most hospitals using similar tools have not.

We argue that every deployed clinical AI system should produce a small, standard set of measures with explicit numerators and denominators, and we propose a minimum dataset for doing so.

2. What changes after deployment

The case for monitoring rests on the observation that the conditions under which a model was validated do not last. Patient populations change, as they did abruptly during the COVID-19 pandemic, when the number of alerts from a widely used sepsis model rose sharply across 24 US hospitals [6]. Documentation, coding and laboratory practices change, sometimes without anyone connecting the change to a model that depends on them [7]. These forms of dataset shift can be anticipated in part when a model is designed [8], but not fully, and their effect on calibration can build up gradually over years [9].

Deployment also changes the thing being predicted. When clinicians act on a model's predictions and prevent the outcome, the model's apparent accuracy falls, and a model retrained on the new data learns that the patients it flagged were at lower risk than they were [10]. The behavior of clinicians changes too. People who use decision support tend to accept its output, especially when checking it is difficult, and to miss what it misses [11]. None of these effects can be seen in development data. All of them can be seen, at least in part, in data from the deployed system.

3. A minimum monitoring dataset

Table 1 sets out the measures we propose. They fall into seven domains, and the guiding principle is that each one is reported as a proportion or rate with its denominator, so that it can be compared over time, between sites and between systems.

DEMO

Table 1. A proposed minimum dataset for monitoring a clinical AI system after deployment. Every measure is reported with its numerator and denominator. Frequencies are suggestions for a system that runs continuously in hospital care and can be adapted to lower-volume uses. EHR, electronic health record.

Domain	Measures	Suggested frequency	Why it matters
Exposure	Encounters eligible, encounters scored and the share scored; share of scores computed with any input missing	Monthly	Shows silent failures, such as broken interfaces or changed codes, that leave patients unscored
Output	Distribution of scores; alerts per 100 patient-days or per encounter, by unit	Weekly	Detects shifts in input data and shows the workload the system creates
Use	Share of alerts opened and acknowledged; overrides, with reasons where recorded; time from alert to the first related order	Monthly	Shows whether the system is used as intended and where clinicians disagree with it
Performance	Discrimination and calibration (observed and expected events) against the target outcome, with the lag needed for outcomes to occur	Quarterly	Detects loss of accuracy and miscalibration; must be read with the model's effect on outcomes in mind
Equity	Exposure, output, use and performance by age, sex, race and ethnicity, preferred language, payer and site	Quarterly	Detects groups that are scored less often, alerted more or less, or predicted less accurately
Safety	Incident reports and near misses linked to the system; cases reviewed and their findings	Continuous, summarized monthly	Captures harms that aggregate statistics miss
Change	Model version; changes to inputs, data mappings, thresholds or display; EHR upgrades and new care pathways	At each change	Allows changes in the other measures to be attributed to their cause

Three features of the dataset need explanation. First, it begins with exposure, not accuracy. A model that silently stops scoring a fifth of patients because an interface changed is a safety problem whatever its area under the curve. Second, it includes use. A system that is accurate but ignored has no effect, and a system that is followed without question can amplify its errors; the proportion of alerts acknowledged and the reasons for overrides are the cheapest available signals of either. Third, performance is reported with its lag and its limits. Outcomes such as sepsis or deterioration are known only after days, and once a model influences care, observed outcomes in flagged patients no longer show what would have happened without it [10]. Monitoring should therefore track calibration in groups that the model did not flag, compare periods before and after major changes, and, where possible, hold out a small random sample from alerting for evaluation. Frameworks for the continual monitoring and updating of clinical AI set out the statistical options for each of these designs [12].

The dataset is deliberately small. Each measure can be computed from data that a system already produces: the scoring log, the alert log, the electronic health record and the incident reporting system. Model facts labels and model cards describe how a system was built and validated [13,14]; the monitoring dataset describes what happens to it afterward, and the two should be published together.

4. Who should collect and report it

DEMO

Responsibility should follow the ability to act. The deploying health system holds the data and controls the response, so it should compute the measures and review them through the same governance that oversees other patient safety programs. Developers, including vendors, should provide the logs and tools that make the measures easy to compute, and should be expected to receive them, because a developer that sees performance data from many sites can detect problems that no single site can.

Regulators are moving toward requirements of this kind but have not yet defined them. In the United States, the Food and Drug Administration's approach to AI-enabled devices that change over time relies on predetermined change control plans, agreed before marketing, that specify what a manufacturer may modify and how it will verify the result [15]. In the European Union, the AI Act (Regulation (EU) 2024/1689) classifies most medical AI as high risk and requires post-market monitoring, although how its obligations will interact with those of the Medical Device Regulation (Regulation (EU) 2017/745) remains unclear [16]. A common minimum dataset would give these requirements a concrete form and allow results to be compared. Much clinical AI, including many models built by health systems for their own use, falls outside device regulation altogether, which makes voluntary standards more important, not less.

Smaller hospitals face a particular problem. A community hospital may see too few events in a quarter to estimate calibration with any precision, and it is unlikely to employ a data scientist. Pooling the measures across hospitals that use the same system, through the developer or through a regional network, would give each site a reference against which to judge its own figures, and would detect problems that affect all sites at once, such as a faulty software update.

Some of the dataset should also be public. Aggregate exposure, alert rates and calibration, published by site at regular intervals, would let clinicians and patients judge a system on its record rather than on its marketing, and would give researchers the denominators needed to interpret early clinical evaluations, which reporting guidelines such as DECIDE-AI already expect to describe how a system was used [17].

5. Objections

The first objection is burden. Many hospitals lack the staff to analyze model performance. But most of the measures in Table 1 are counts and proportions that can be produced automatically, and the cost of not knowing includes the clinician time consumed by inaccurate alerts and the harm from missed cases. Shared tools, provided by developers or by networks of health systems, would reduce the burden further.

The second is commercial confidentiality. Developers may be reluctant to see site-level performance published. Yet the measures describe how a product performs in a customer's hospital on that hospital's patients, and the customer has a legitimate interest in knowing it. Purchase contracts can require access to the data needed for monitoring.

The third is that the measures can mislead. Falling performance may reflect successful prevention rather than failure, and differences between groups may be due to chance. This is a reason to report the measures with their denominators, their uncertainty and the change log that allows them to be interpreted, not a reason to avoid collecting them.

The fourth is that measures become targets. If a health system is judged on its alert acknowledgment rate, clinicians may learn to acknowledge alerts without reading them. The dataset is meant for learning

DEMO

about a system, not for ranking units or individuals, and its governance should say so. Measures of use are most informative when read alongside measures of performance and safety, which are harder to game.

6. Conclusion

Clinical AI systems are judged at the point of purchase on evidence that is rarely updated afterward. A small set of measures with explicit denominators, covering whether patients were scored, what the system said, whether anyone acted, whether it was right, for whom, with what harms and after which changes, would let health systems, developers and regulators see how these systems behave in practice. The data already exist in the logs of every deployed system. The step that is missing is to agree what to compute and to compute it.

Declarations

Ethics approval and consent to participate	Not applicable.
Data availability	No new data were generated or analyzed for this article. Code and model availability: not applicable.
Use of generative AI	Not applicable to the fictional study. This demonstration article was drafted with the help of a generative AI tool so that the journal's article design could be previewed; it reports no real research.
Funding	This work received no specific funding.
Competing interests	The authors declare no competing interests.
Author contributions (CRediT)	Conceptualization: NBA, COL; Writing, original draft: NBA; Writing, review and editing: RI, COL. All authors approved the final version and agree to be accountable for the work.
Acknowledgments	This is a demonstration article created to preview the journal's design. The study, its data and its authors are fictional.
Peer review	Not peer reviewed: this is a demo article. A published article names its review model here (double-blind, external reviewers) and its handling editor.

References

1. Benjamens S, Dhunnoo P, Meskó B. The state of artificial intelligence-based FDA-approved medical devices and algorithms: an online database. NPJ Digit Med. 2020;3(1):118. doi:10.1038/s41746-020-00324-0

2. Muehlematter UJ, Daniore P, Vokinger KN. Approval of artificial intelligence and machine learning-based medical devices in the USA and Europe (2015-20): a comparative analysis. Lancet Digit Health. 2021;3(3):e195-e203. doi:10.1016/S2589-7500(20)30292-2

DEMO

3. Wu E, Wu K, Daneshjou R, Ouyang D, Ho DE, Zou J. How medical AI devices are evaluated: limitations and recommendations from an analysis of FDA approvals. Nat Med. 2021;27(4):582-584. doi:10.1038/s41591-021-01312-x

4. Wong A, Otles E, Donnelly JP, et al. External validation of a widely implemented proprietary sepsis prediction model in hospitalized patients. JAMA Intern Med. 2021;181(8):1065-1070. doi:10.1001/jamainternmed.2021.2626

5. Habib AR, Lin AL, Grant RW. The Epic Sepsis Model falls short: the importance of external validation. JAMA Intern Med. 2021;181(8):1040-1041. doi:10.1001/jamainternmed.2021.3333

6. Wong A, Cao J, Lyons PG, et al. Quantification of sepsis model alerts in 24 US hospitals before and during the COVID-19 pandemic. JAMA Netw Open. 2021;4(11):e2135286. doi:10.1001/jamanetworkopen.2021.35286

7. Finlayson SG, Subbaswamy A, Singh K, et al. The clinician and dataset shift in artificial intelligence. N Engl J Med. 2021;385(3):283-286. doi:10.1056/NEJMc2104626

8. Subbaswamy A, Saria S. From development to deployment: dataset shift, causality, and shift-stable models in health AI. Biostatistics. 2020;21(2):345-352. doi:10.1093/biostatistics/kxz041

9. Davis SE, Lasko TA, Chen G, Siew ED, Matheny ME. Calibration drift in regression and machine learning models for acute kidney injury. J Am Med Inform Assoc. 2017;24(6):1052-1061. doi:10.1093/jamia/ocx030

10. Lenert MC, Matheny ME, Walsh CG. Prognostic models will be victims of their own success, unless... J Am Med Inform Assoc. 2019;26(12):1645-1650. doi:10.1093/jamia/ocz145

11. Lyell D, Coiera E. Automation bias and verification complexity: a systematic review. J Am Med Inform Assoc. 2017;24(2):423-431. doi:10.1093/jamia/ocw105

12. Feng J, Phillips RV, Malenica I, et al. Clinical artificial intelligence quality improvement: towards continual monitoring and updating of AI algorithms in healthcare. NPJ Digit Med. 2022;5(1):66. doi:10.1038/s41746-022-00611-y

13. Sendak MP, Gao M, Brajer N, Balu S. Presenting machine learning model information to clinical end users with model facts labels. NPJ Digit Med. 2020;3(1):41. doi:10.1038/s41746-020-0253-3

14. Mitchell M, Wu S, Zaldivar A, et al. Model cards for model reporting. In: Proceedings of the Conference on Fairness, Accountability, and Transparency. Association for Computing Machinery; 2019:220-229. doi:10.1145/3287560.3287596

15. Vokinger KN, Feuerriegel S, Kesselheim AS. Continual learning in medical devices: FDA's action plan and beyond. Lancet Digit Health. 2021;3(6):e337-e338. doi:10.1016/S2589-7500(21)00076-5

16. Gilbert S. The EU passes the AI Act and its implications for digital medicine are unclear. NPJ Digit Med. 2024;7(1):135. doi:10.1038/s41746-024-01116-6

17. Vasey B, Nagendran M, Campbell B, et al. Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. Nat Med. 2022;28(5):924-933. doi:10.1038/s41591-022-01772-9

Open access. Copyright 2026 the authors. This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (https://creativecommons.org/licenses/by/4.0/), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, provided the original authors and source are credited.
Published by Eldenhall Sciences Inc., 117 South Lexington Street, Suite 100, Harrisonville, MO 64701, United States. editorial@eldenhall.org.
Online version: https://eldenhall.org/articles/post-deployment-monitoring-clinical-ai-minimum-dataset/

DEMO