HHS Public Access

Author manuscript

NEJM AI. Author manuscript; available in PMC 2026 January 29.

Published in final edited form as:

NEJM AI. 2025 December ; 2(12): . doi:10.1056/aip2500705.

Evaluating Translational AI: A Two-Way Moving Target Problem

Richard K. Leuchter, M.D.[1,2], **William B. Turner, B.A.**[1], **David Ouyang, M.D.**[3]

[1]Department of Medicine, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, CA, USA

[2]Greater Los Angeles Veterans Affairs Healthcare, Los Angeles, CA, USA

[3]Department of Cardiology, Cedars-Sinai Medical Center, Smidt Heart Institute, Los Angeles, CA, USA

Abstract

Predictive artificial intelligence models are being deployed across health systems with dangerously inconsistent oversight, creating two critical gaps: a compliance gap, where clinical tools that likely qualify as software as a medical device are implemented without seeking U.S. Food and Drug Administration authorization; and a regulatory gap, where administrative and operational models are deployed without any external review despite their potential to influence care and widen disparities. Given that comprehensive U.S. Food and Drug Administration oversight of all such models is infeasible, the de facto onus of ensuring their safety and efficacy falls on the implementing institutions. However, this imperative for self-governance is undermined by a fundamental and previously unarticulated two-way moving target problem: (1) prior to implementation, concurrent-intervention confounding moves the target as practice and operational changes shift the outcome during the time it takes to develop models; and (2) after implementation, action-induced outcome bias moves the target again when prediction-triggered interventions alter or censor the outcome. Together, these pitfalls render traditional evaluation methods inadequate. The authors argue that health systems must adopt a new default standard for implementing any model that predicts patient outcomes or utilization: short-term randomized deployment with a control group. This approach provides a crucial counterfactual for rigorous, independent assessment of model performance and intervention effectiveness. It offers a practical path forward for institutions to ensure that their artificial intelligence tools are safe, effective, and equitable, thereby building a foundation of trust that is worthy of the patients they serve. (Funded by the National Institutes of Health National Heart, Lung, and Blood Institute.)

A Regulatory Problem

In the rush to integrate artificial intelligence (AI) into clinical practice, a powerful and increasingly common class of predictive models is quietly proliferating. These models can

Richard K. Leuchter can be contacted at rleuchter@mednet.ucla.edu or at the Division of General Internal Medicine and Health Services Research, David Geffen School of Medicine, University of California, Los Angeles, 1100 Glendon Avenue, #800, Los Angeles, CA 90024.

Disclosures

Author disclosures are available at ai.nejm.org.

be administrative tools that predict operational metrics, but are often clinical models that skirt the U.S. Food and Drug Administration (FDA)'s software as a medical device (SaMD) regulations. A prominent example is the Epic Sepsis Model, a tool introduced to hundreds of hospitals in the United States to detect a life-threatening condition. It was deployed without seeking FDA authorization[1] and — as later independent evaluations revealed — without undergoing sufficiently rigorous evaluation to ensure its effectiveness.[2,3] The model remains in widespread use today, along with newer third-party models to detect sepsis,[4] without FDA clearance despite a clear statement from the FDA that "software device[s] to aid in the prediction or diagnosis of sepsis" are considered SaMD and subject to regulation.[5] Simultaneously, a large number of AI models designed to optimize billing and risk stratification, which are not subject to SaMD regulations,[6] are being adopted, but these may unintentionally widen disparities in resource allocation.[7] These scenarios reveal two gaps in the approach to AI governance: a compliance gap, where institutions and vendors deploy SaMD models without seeking the requisite FDA authorization; and a regulatory gap, where administrative models are deployed without external oversight despite their ability to influence care.

These gaps apply to a specific category of AI models: translational models that predict patient outcomes or resource utilization (e.g., clinical deterioration, length of stay, readmission, no show) and that inform or trigger clinical or operational actions. Half the challenge with this application of health care AI is that developers of models that assist with diagnosing but do not explicitly drive clinical decisions previously exploited a loophole in SaMD regulation, marketing them as clinical decision support (CDS) tools, which are exempt from FDA regulation per the 21st Century Cures Act.[8] FDA guidance has subsequently closed this loophole with more stringent criteria for qualifying as a CDS tool,[9] yet health systems and third-party vendors continue to implement these models without FDA authorization, using the outdated CDS justification.[4] The other half of the challenge is that nondiagnostic administrative models, such as those predicting length of stay, can be implemented without FDA oversight and are not held to rigorous performance standards.

This dual problem creates a dual imperative of institutional self-governance. First, the sheer volume of AI models for predicting patient outcomes and utilization produced each year would cause a "death by regulation" scenario if the FDA oversaw them all. Instead, the FDA upholds an "enforcement discretion" policy,[10] meaning that the de facto onus of ensuring the safety of these AI models falls on the implementing health system. Second, for administrative/operational models that fall outside the FDA's purview, health systems must ensure that these models do not cause unintended consequences. For models that predict patient outcomes and utilization, a new evaluation framework is needed — one centered on institutional accountability.

An Evaluation Problem

Despite the need for institutional accountability, the unfortunate landscape of these models is plagued by a research–implementation gap in which researched models are rarely implemented, and implemented models are inadequately researched.[11] Peer-reviewed AI models are published almost daily, yet very few make it into a live electronic health

record environment due to substantial implementation challenges.[12] Conversely, as the Epic Sepsis Model demonstrates, implemented models may not be rigorously validated, leading to poor real-world performance that goes undetected until widespread deployment.[2,3] This gap between implementation and evaluation underscores the need for institutional self-governance, but models that predict patient outcomes or utilization and inform an action are subject to a two-way moving target problem comprising two unique evaluation challenges (Fig. 1).

The first challenge is concurrent-intervention confounding, which makes model predictions inaccurate from the moment of implementation, even when performance metrics were excellent during model validation. During the months it takes to train a model on retrospective data (e.g., length of stay), health systems often implement concurrent interventions to solve the same problem (e.g., new discharge initiatives). This describes the first moving target: by the time the AI model is deployed, the underlying concept it was trained to predict has fundamentally changed (i.e., concept drift).[13] Nonrandom rollout timing, cointerventions, secular trends, and seasonality add to this moving target by further shifting the data-generation process in the time between model development and real-world use. Techniques such as shadow mode,[14] in which the model is run silently on live data, help to overcome this challenge by allowing for contemporaneous retraining, but further delay implementation.

The second challenge is action-induced outcome bias: the target moves again when a prediction triggers an intervention that alters the outcome, making an accurate forecast appear wrong.[15] If a high predicted length of stay prompts early discharge planning and the stay is shortened, the observed outcome no longer reflects what would have occurred without the prediction, and the original forecast can appear wrong despite being prescient. This feedback introduces exposure-induced bias and intervention-induced censoring; prospective calibration, discrimination, and pre–post comparisons conflate model performance with intervention effect.[16,17] In addition, regression to the mean among flagged “outliers” can spuriously magnify before-and-after improvements in exposed patients.[18] Consequently, before-and-after evaluations cannot identify causal impact because the necessary counterfactual (what would have happened absent exposure to the prediction or the triggered action) is missing.[16]

These pitfalls are largely specific to outcome/utilization prediction models that trigger actions and generally do not arise for tasks with fixed reference standards (e.g., radiology interpretation or ambient scribe transcription), where accuracy is judged against a stable ground truth rather than a counterfactual altered by the prediction.

A Path Forward: Institutional Rigor through Randomized Implementation

To fulfill the dual imperative of self-governance while also overcoming the two types of moving target problems, we propose that health systems publicly commit to a default short-term randomized deployment with a concurrent control for any model that predicts patient outcomes or utilization. A contemporaneous control that is unaffected by alerts or downstream actions lets teams detect concurrent-intervention confounding and recalibrate

in real time using prospective data, rather than relying on before–and–after checks that are vulnerable to timing effects and regression to the mean. It also preserves the counterfactual necessary to assess model calibration and discrimination without contamination from action-induced outcome bias.

This proposal is operationally feasible through the following practical randomized deployment designs: (1) alert-suppression randomization, where model predictions are made for all, but alerts are randomly shown versus suppressed; (2) stepped-wedge cluster rollout when individual-level randomization is impractical; or (3) two-stage designs that decouple prediction from action (e.g., randomize alert visibility, then, among visible alerts, randomize the action). These designs cleanly separate model performance from intervention effects, enable early drift detection, and provide credible estimates of safety and unintended harms. A minimum standard should include preregistration, a brief statistical analysis plan (including a justification for the randomization duration), and prespecified equity analyses (e.g., stratified by sex and/or insurance coverage), with transparent reporting of results.

Conclusion

Many, perhaps most, internally developed models predicting patient outcomes or utilization may not require full FDA review. However, an absence of oversight cannot justify an absence of rigor. Health systems must hold themselves accountable. Doing so requires confronting a two-way moving target: the outcome shifts both before deployment as practices evolve and after deployment in response to model-informed interventions. Adopting randomized implementation as the default standard for this growing class of AI models ensures that tools are not only powerful in theory but proven safe, effective, and equitable in practice. This is the only way to build a foundation of trust that is worthy of the patients we serve.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Dr. Leuchter was funded by a grant from the National Institutes of Health National Heart, Lung, and Blood Institute (grant no. 1K38 HL164955-01). The funders had no role in considering the study design or in the collection, analysis, or interpretation of data; writing of the report; or the decision to submit the article for publication.

References

1. Zhang J, Mattie H, Shuaib H, Hensman T, Teo JT, Celi LA. Addressing the “elephant in the room” of AI clinical decision support through organisation-level regulation. PLoS Digit Health 2022;1:e0000111. DOI: 10.1371/journal.pdig.0000111. [PubMed: 36812576]
2. Kamran F, Tjandra D, Heiler A, et al. Evaluation of sepsis prediction models before onset of treatment. NEJM AI 2024;1(3). DOI: 10.1056/AIoa2300032.
3. 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;181:1065–1070. DOI: 10.1001/jamainternmed.2021.2626. [PubMed: 34152373]

4. Patrick K. Early warning for sepsis from Bayesian health — advances of AI in clinical care continue. September 8, 2022 (https://www.signifyresearch.net/insights/early-warning-for-sepsis-from-bayesian-health-advances-of-ai-in-clinical-care-continue/).
5. U.S. Food and Drug Administration. Product classification. October 13, 2025 (https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfPCD/classification.cfm?id=SAK).
6. Sahni NR, Carrus B. Artificial intelligence in U.S. health care delivery. N Engl J Med 2023;389:348–358. DOI: 10.1056/NEJMra2204673. [PubMed: 37494486]
7. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. Science 2019;366:447–453. DOI: 10.1126/science.aax2342. [PubMed: 31649194]
8. U.S. Food and Drug Administration. Changes to existing medical software policies resulting from section 3060 of the 21st century cures act. September 27, 2019 (https://www.fda.gov/media/109622/download).
9. U.S. Food and Drug Administration. Clinical decision support software. September 28, 2022 (https://www.fda.gov/media/109618/download).
10. U.S. Food and Drug Administration. Step 7: does the device software functions (DSF) and mobile medical applications (MMA) guidance apply? December 12, 2022 (https://www.fda.gov/medical-devices/digital-health-center-excellence/step-7-does-device-software-functions-dsf-and-mobile-medical-applications-mma-guidance-apply).
11. Sing K. Tackling the health AI paradox. In: Wu S, ed. Proceedings of Medical Imaging 2025: Imaging Informatics. SPIE Medical Imaging, San Diego, CA: 2025:1341102. DOI: 10.1117/12.3055542.
12. Rahimi AK, Pienaar O, Ghadimi M, et al. Implementing AI in hospitals to achieve a learning health system: systematic review of current enablers and barriers. J Med Internet Res 2024;26:e49655. DOI: 10.2196/49655. [PubMed: 39094106]
13. Lima M, Neto M, Filho TS, Roberta RA. Learning under concept drift for regression — a systematic literature review. IEEE Access 2022;10:45410–45429. DOI: 10.1109/ACCESS.2022.3169785.
14. Christopher GS. Deploying machine learning models in shadow mode. April 12, 2020 (https://christophergs.com/machinelearning/2019/03/30/deploying-machine-learning-applications-in-shadow-mode).
15. Logan Ellis H, Palmer E, Teo JT, Whyte M, Rockwood K, Ibrahim Z. The early warning paradox. NPJ Digit Med 2025;8:1–2. DOI: 10.1038/s41746-024-01408-x. [PubMed: 39747648]
16. Hernán MA, Robins JM. Causal inference: what if. Boca Raton: Chapman & Hall/CRC. 2020 (https://www.hsph.harvard.edu/miguel-).
17. Robins JM, Hernán MÁ, Brumback B. Marginal structural models and causal inference in epidemiology. Epidemiology 2000;11:550–560. DOI: 10.1097/00001648-200009000-00011. [PubMed: 10955408]
18. Barnett AG, van der Pols JC, Dobson AJ. Regression to the mean: what it is and how to deal with it. Int J Epidemiol 2005;34:215–220. DOI: 10.1093/ije/dyh299. [PubMed: 15333621]

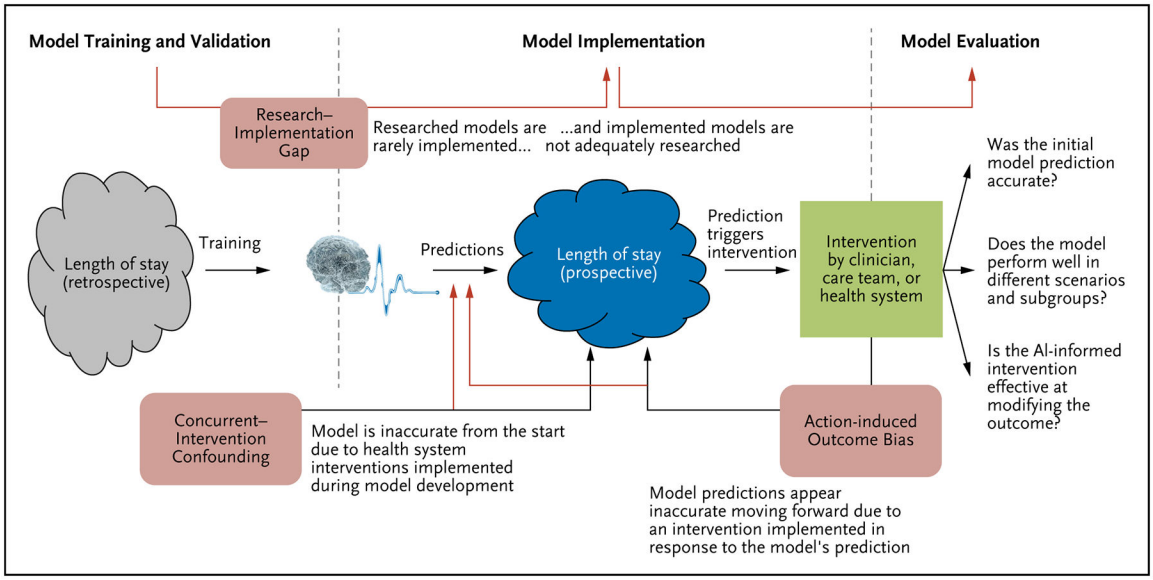

Figure 1.

Summary of the Two-Way Moving Target Problem That Affects Translational Artificial Intelligence Models Predicting Patient Outcomes and Utilization, Using a Model Predicting Length of Stay as an Example.

The evaluation of artificial intelligence models that predict outcomes like length of stay is complicated by two key pitfalls. The first is concurrent-intervention confounding: the model's predictions are inaccurate on implementation because concurrent interventions have altered the clinical environment from what was captured in the retrospective training data (i.e., data drift and concept drift). The second is action-induced outcome bias: a successful intervention, triggered by an accurate prediction, alters or censors the outcome, making the prediction appear inaccurate and confounding the assessment of model and intervention efficacy (i.e., prediction decay). AI denotes artificial intelligence.