npj | digital medicine

News & Views

Published in partnership with Seoul National University Bundang Hospital

https://doi.org/10.1038/s41746-024-01242-1

Ethical debates amidst flawed healthcare artificial intelligence metrics

Jack Gallifant, Danielle S. Bitterman, Leo Anthony Celi, Judy W. Gichoya, Joao Matos, Liam G. McCoy & Robin L. Pierce

Healthcare AI faces an ethical dilemma between selective and equitable deployment, exacerbated by flawed performance metrics. These metrics inadequately capture real-world complexities and biases, leading to premature assertions of effectiveness. Improved evaluation practices, including continuous monitoring and silent evaluation periods, are crucial. To address these fundamental shortcomings, a paradigm shift in AI assessment is needed, prioritizing actual patient outcomes over conventional benchmarking.

Artificial intelligence (AI) is poised to bridge the deployment gap with increasing capabilities for remote patient monitoring, handling of diverse time series datasets, and progression toward the promise of precision medicine. This proximity also underscores the urgency of confronting the translational risks accompanying this technological evolution and maximizing alignment with fundamental principles of ethical, equitable, and effective deployment. The recent work by Goetz et al. surfaces a critical issue at the intersection of technology and healthcare ethics: the challenge of generalization and fairness in health AI applications[1]. This is a complex issue where equal performance across subgroups can be at odds with overall performance metrics[2].

Specifically, it highlights one potential avenue to navigate variation in model performance among subgroups based on the concept of "selective deployment"[3]. This strategy asserts that limiting the deployment of the technology to the subgroup in which it works well facilitates benefits for those subpopulations. The alternative is not to deploy the technology in the optimal performance group but instead adopt a standard of equity in the performance overall to achieve parity among subgroups, what might be termed "equitable deployment". Some view this as a requirement to "level down" performance for the sake of equity, a view that is not unique to AI or healthcare and is the subject of a broader ethical debate[4–6]. Proponents of equitable deployment would counter: Can a commitment to fairness justify not deploying a technology that is likely to be effective but only for a specific subpopulation?

Discussions around selective deployment do not take place in a vacuum and must be had with an awareness of the broader context of the attributes at hand and the sociopolitical context of healthcare. Healthcare inherently involves unequal distribution of resources—however it is not the unequal allocation per se, but rather the underlying basis for such allocation that demands scrutiny. While selective deployment may, by a first-pass utilitarian calculus, lead to more patient benefit than withholding a model until equitable performance can be achieved, the second-order impacts are more complicated.

This article seeks to offer reflections on not only the ethical tensions between selective and equitable deployment but also on the myriad barriers that render real-world equity much more complicated than in silico validations imply. Without significant improvements in health AI metrics and evaluation practices, these debates will continue to take place in an environment of insufficient information to determine a policy's ultimate impact.

Informational challenges in selective deployment

One formulation of selective deployment posits that models should be confined to subpopulations with robustly validated efficacy and safety. This is framed as a prudent exercise in risk mitigation. After all, deploying unvalidated models in clinical settings without demonstrable efficacy invites a litany of ethical and legal complications, potentially compromising patient care and contravening medical ethics principles. This concern intersects with a much broader challenge in AI in medicine—what does it mean for a model to be "validated" for *any* subpopulation, let alone the population at large?

At its core, a model's objective should be to elucidate and reflect the actual causal factors underlying patient conditions, aiming to enhance, predict, and ultimately improve patient outcomes. Nonetheless, this pursuit must address the inevitable presence of 'bias exhaust'—residual biases not directly related to biological determinants but rather emerge from systemic discrepancies in care, data processing, and various operational protocols. Identifying models that most accurately capture these causal relationships represents the zenith of machine learning applications in healthcare. However, this objective encounters substantial hurdles given the often incomplete understanding of causal pathways in medicine and the fact that training data is a mere abstraction of complex biological, clinical, and sociocultural realities faced by patients and healthcare providers.

The ability of models to represent these pathways dependably is further abstracted via processing, selection, and imputation. When combined with modern high-powered architectures, such as neural networks, which are particularly susceptible to overfitting, we can compound errors relying on spurious correlations and superficial features that boost apparent performance in development but may not accurately represent the underlying causal mechanisms of health outcomes at the bedside. Critically, therefore, a model's ultimate veracity must be demonstrated *not* during its development phase but upon its deployment in real-world clinical settings.

Given that the initial development period represents the most favorable context for claimed performance, it is particularly concerning if there is already an assertion of ineffectiveness in other groups at the development stage. Further, even optimal performance in a particular subgroup during initial benchmarking represents, at best, a premature and unvalidated

Box 1 | Strategies for achieving more than in silico accuracy

- *Access to high-resolution real-world data*: Provide developers with diverse, comprehensive clinical datasets to train models on actual patient populations and scenarios.
- *Systematic evaluation pipelines*: Implement robust data pipelines to continuously assess model performance and patient outcomes across various demographic and clinical subgroups.
- *Data shift monitoring*: Develop dashboards to track changes in data distributions over time, alerting to potential model drift and ensuring ongoing relevance.
- *Accountability frameworks*: Establish clear responsibilities and oversight mechanisms for all stakeholders involved in the AI model lifecycle, from development to deployment.
- *Mandatory silent evaluation periods*: Require a phase of background performance assessment in real clinical settings before active deployment, focusing on safety, efficacy, and equity.
- *Multidisciplinary collaboration*: Engage healthcare professionals, patients, and social scientists to define legitimate subgroup differences and ensure culturally competent AI systems.
- *Iterative refinement process*: Implement a feedback loop for continuous model improvement based on real-world performance data and stakeholder input.
- *Transparency in reporting*: Mandate clear documentation of model limitations, potential biases, and performance variations across different populations.

assertion of effectiveness. In many contexts, the claimed subgroup benefits that provide the foundation for justifying selective deployment despite equity costs may be far from as great as they initially seem.

Contextual relevance and irrelevance

The arguments for selective deployment raise an adjacent problem surrounding evaluating real-world differences in model outcomes: distinguishing between which factors should reasonably influence results and which factors reflect impermissible bias. Bias, a generally elusive term often regarded as pejorative, suggests a model's failure to accurately mirror the multifaceted realities of patient conditions and health determinants[7]. While certain factors may be felt to play a legitimate causal role, others simply embed and reflect societal biases that should not be reproduced.

Assessing when differential performance across groups is acceptable depends in large part upon distinguishing between these acceptable and unacceptable drivers of performance. This may sometimes be predicated on legitimate variability of medical conditions, their manifestations, and biological makeup. For instance, the clinical presentation and diagnostic markers may inherently differ between sexes, necessitating distinct modeling approaches to predict outcomes accurately, such as the breast cancer example in Goetz et al. However, when a model's performance is disproportionately influenced by factors that should not significantly affect the diagnostic or prognostic accuracy—such as a patient's race in contexts where it should not bear relevance[8]—this disparity signals a deviation from the pursuit of accurate causal representation. However, part of the challenge in this field lies in the complexity of identifying relevant subgroups and how to handle intersectionality, particularly when definitions are based on social and legal constructs, such as race, rather than biological differences.

In addition to embedding problematic societal biases, predictive factors that reflect social contingencies rather than legitimate biological causality may be particularly brittle when applied within the complex and unpredictable environment of clinical practice. These factors may lead to a degradation of model performance upon deployment in the real world, underscoring a fundamental misalignment between the model's training and the realities it is intended to navigate[9].

Towards better metrics for performance

This discussion underscores a critical paradigm shift: the evaluation of machine learning models in healthcare must extend beyond conventional metrics of accuracy attained during the benchmarking phase (Box 1). High accuracy in a controlled test environment may not translate to effectiveness at the bedside, where the intricacies of patient care unfold in real-time. Once the model is deployed, the same metrics used for training may no longer be applicable for post-deployment monitoring, as the model itself can modify clinicians' actions and patients' trajectories. The ultimate measure of a model's value lies in its ability to improve patient outcomes in authentic deployment contexts.

Nuanced evaluation of complex downstream metrics at the time of model deployment is required, yet such data is, by its nature, unavailable to AI teams at the time of initial development. Therefore iterative silent evaluation can help to bridge this gap, wherein models' performance is evaluated in the background of ongoing clinical activity without yet impacting patient care[10,11]. This is but a portion of a robust framework for continuous evaluation and accountability that encompasses the entire lifecycle of model development, from data collection and curation to deployment, including a rigorous regimen of post-market surveillance to monitor a model's performance in live healthcare settings. Such mechanisms must act to ensure that disparities do not widen, models receive timely recalibration and error detection processes, ensure models adapt to the evolving landscape of healthcare data and patient demographics. This is particularly relevant given the significant subpopulation shifts these models will encounter and the rate of change in clinical knowledge being deployed by physicians in practice.

Moreover, contemporary debates regarding selective and equitable deployment are driven by the same flawed sets of performance metrics that the field at large must work to improve. Without a clearer understanding of the downstream impact, ethical discussions at the time of development are blind to essential information underpinning them. Regardless of the stance one takes in the debate between selective and equitable deployment, improvements in evaluation practices are essential to ensuring that the debate is well-informed and connected to actual impact.

Jack Gallifant[1,2], Danielle S. Bitterman[3,4,5], Leo Anthony Celi[1,6,7] ✉, Judy W. Gichoya[8], Joao Matos[1,9,10], Liam G. McCoy[11] & Robin L. Pierce[12]

[1]Laboratory for Computational Physiology, Massachusetts Institute of Technology, Cambridge, MA, USA. [2]Department of Critical Care, Guy's and St Thomas' NHS Foundation Trust, London, UK. [3]Artificial Intelligence in Medicine (AIM) Program, Mass General Brigham, Harvard Medical School, Boston, MA, USA. [4]Department of Radiation Oncology, Brigham and Women's Hospital/Dana-Farber Cancer Institute, Boston, MA, USA.

[5]Computational Health Informatics Program, Boston Children's Hospital, Harvard Medical School, Boston, MA, USA. [6]Division of Pulmonary, Critical Care and Sleep Medicine, Beth Israel Deaconess Medical Center, Boston, MA, USA. [7]Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA, USA. [8]Department of Radiology, Emory University School of Medicine, Georgia, USA. [9]Faculty of Engineering, University of Porto, Porto, Portugal. [10]Institute for Systems and Computer Engineering, Technology and Science, Porto, Portugal. [11]Faculty of Medicine and Dentistry, University of Alberta, Edmonton, Canada. [12]The Law School, Faculty of Humanities, Arts, and Social Sciences, University of Exeter, Exeter, UK. ✉e-mail: lceli@mit.edu

Received: 15 May 2024; Accepted: 29 August 2024;
Published online: 11 September 2024

References

1. Goetz, L., Seedat, N., Vandersluis, R., & van der Schaar, M. Generalization—a key challenge for responsible AI in patient-facing clinical applications. *Npj Digit. Med.* **7**, 1–4 (2024).
2. D'Amour, A. et al. Underspecification presents challenges for credibility in modern machine learning. *J. Mach. Learn. Res.* **23**, 226:10237–226:10297 (2022).
3. Vandersluis, R. & Savulescu, J. The selective deployment of AI in healthcare. *Bioethics* **38**, 391–400 (2024).
4. Seidman, L. M. The Ratchet wreck: equality's leveling down problem. *Ky. Law J.* **110**, 59–106 (2021).
5. Thomas, T. A. Leveling down gender equality. *Harv. J. Law Gend.* **42**, 177–218 (2019).
6. Sessions v. Morales-Santana, 137 S. Ct. 1678. https://supreme.justia.com/cases/federal/us/582/15-1191 (2017).
7. Abràmoff, M. D. et al. Considerations for addressing bias in artificial intelligence for health equity. *Npj Digit. Med.* **6**, 1–7 (2023).
8. Basu, A. Use of Race in Clinical Algorithms—PMC. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10219586/.
9. Finlayson Samuel, G. et al. The clinician and dataset shift in artificial intelligence. *N. Engl. J. Med.* **385**, 283–286 (2021).
10. Nestor, B. et al. Preparing a clinical support model for silent mode in general internal medicine. In *Proc. of the 5th Machine Learning for Healthcare Conference* (PMLR) (eds Doshi-Velez et al.), pp. 950–972 (2020).
11. Kwong, J. C. C. et al. The silent trial-the bridge between bench-to-bedside clinical AI applications. *Front. Digit. Health* **4**, 929508 (2022).

Acknowledgements

J.G. is funded by the National Institute of Health through DS-I Africa U54 TW012043-01 and Bridge2AI OT2OD032701. D.B. declares funding from the NIH (NIH-USA U54CA274516-01A1 and R01CA294033-01). L.A.C. is funded by the National Institute of Health through R01 EB017205, DS-I Africa U54TW012043-01, Bridge2AI OT2OD032701, and the National Science Foundation through ITEST #2148451. JWG is a 2022 Robert Wood Johnson Foundation Harold Amos Medical Faculty Development Program and declares support from the RSNA Health Disparities grant (#EIHD2204), Lacuna Fund (#67), Gordon and Betty Moore Foundation, and NIH (NIBIB) MIDRC grant under contracts 75N92020C00008 and 75N92020C00021.

Author contributions

J.G., J.M., and L.A.C. drafted the initial manuscript and edited subsequent versions. D.S.B., J.W.G., L.G.M., and R.L.P. provided critical revisions, feedback, and edits to the manuscript drafts. All authors reviewed and approved the final version.

Competing interests

L.A.C.: Editor in Chief of PLoS Digital Health; D.B.: Associate Editor of Radiation Oncology, HemOnc.org (no financial compensation, unrelated to this work). All other authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to Leo Anthony Celi.

Reprints and permissions information is available at http://www.nature.com/reprints