

Viewpoint

Algorithmic blind spots: bias, opacity and the clinical translation gap in artificial intelligence-driven personalized medicine

Atif Adnan

Institute of Life Sciences, Shenyang Medical College, Shenyang, P.R. China; Liaoning Province Key Laboratory for Phenomics of Human Ethnic Specificity and Critical Illness, Shenyang Medical College, Shenyang, P.R. China; Liaoning Provincial Belt and Road Joint Laboratory for China-Pakistan Human Phenomics Research, Shenyang Medical College, Shenyang, P.R. China

Correspondence: mirzaatifadnan@gmail.com



Citation: Adnan A. Algorithmic blind spots: bias, opacity and the clinical translation gap in artificial intelligence-driven personalized medicine. *J Basic Clin Med Sci.* 2026;5(1):54-58.

Received: 08 October 2025

Revised: 15 March 2026

Accepted: 21 March 2026

Published: 30 March 2026



Copyright: © 2026 The Author. This is an open access article distributed under the terms of the [Creative Commons Attribution \(CC BY\) License](https://creativecommons.org/licenses/by/4.0/).

Extract

The use of machine learning has changed how molecular and clinical data are used to detect disease, forecast prognosis and select treatment, yet a gap remains between reported performance and clinical use. Few published models have been tested prospectively, and several that have reached practice perform far worse than expected. From this perspective, I argue that clinical translation should be seen as a continuum from data to decision. Three weaknesses explain much of the gap: models whose outputs cannot be checked against biology, data that are unrepresentative or inconsistently recorded, and the absence of shared standards for evaluation and oversight. I propose a five-stage framework that links representative, high-quality data; clinically interpretable models; external and prospective validation; workflow integration; and postdeployment monitoring and governance. Low- and middle-income countries need locally governed data and regional validation networks rather than imported tools. The goal is safer, more equitable clinical decisions, not better prediction alone.

Keywords

Artificial intelligence; Machine learning; Precision medicine; Molecular diagnostic techniques; Healthcare disparities; Validation studies as topic; Bias

Integrating machine learning and artificial intelligence (AI) into molecular diagnostics has changed how disease detection, prognostic forecasting and treatment selection are approached [1]. Models that draw on genomic, transcriptomic, imaging and electronic health record (EHR) data promise care tailored to the individual patient. I refer to this goal as personalized medicine and use the term “precision medicine” synonymously. Published reports of predictive performance are often impressive, but the record in clinical practice is far weaker. A systematic review of deep learning studies in medical imaging revealed that of 81 nonrandomized studies comparing algorithms with clinicians, only nine were prospective, and only six were tested in a real clinical setting; nevertheless, the performance of 61 was at least comparable to that of clinicians [2]. A later review revealed only 41 randomized trials of machine learning interventions in health care; none adhered fully to the Consolidated Standards of Reporting Trials–Artificial Intelligence (CONSORT-AI) reporting guidelines, and approximately half were conducted at a single site [3]. A 2024 scoping review of 86 randomized trials likewise revealed a predominance of single-center studies with little demographic reporting [4]. The evidence that these tools help patients lags well behind the enthusiasm for them [5].

I argue that clinical translation is best understood as a continuum running from data to decision rather than as a single validation step. Even a model with strong reported performance can fail if its input data arrives late or are incomplete, if it fits poorly into the clinical workflow, or if its predictions are not linked to a clear clinical action [5]. Weaknesses at three points along this continuum explain much of the gap: models whose outputs cannot be checked against biology, data that are neither representative nor consistently recorded, and the lack of shared standards for evaluation and oversight. In response, I propose that clinical AI should pass through five linked stages: representative, high-quality data; a clinically interpretable model; external and prospective validation; integration into clinical workflow and decisions; and postdeployment monitoring and governance, with findings from monitoring feeding back into each earlier stage.

The first weakness is the distance between what an algorithm concludes and what is biologically plausible. Many deep learning systems are opaque, mapping inputs to clinical outcomes through statistical association with no account of the biology underneath. When a model latches onto a surrogate variable or an artifact of how the data were collected, rather than a true driver of disease, its performance tends to collapse in external testing. Zech and colleagues showed this with a pneumonia model trained on chest radiographs. Because the prevalence of pneumonia differed sharply across hospitals, the model performed worse on data from other hospital systems than on internal data [6]. During the coronavirus disease 2019 (COVID-19) pandemic, a living systematic review of 232 prediction models revealed that every one had a high or unclear risk of bias [7]. An external validation of a proprietary sepsis model used in hundreds of hospitals in the United States revealed an area under the receiver operating characteristic curve (AUROC) of 0.63; the model missed 67% of patients who developed sepsis while generating alerts for 18% of all hospitalized patients [8].

It helps to separate three properties that are often conflated. Predictive validity means that a model forecasts outcomes reliably in the population where it will be used. Clinical interpretability means that the intended user can understand enough about how an output was produced to judge whether to act on it. Mechanistic or causal validity means that the features driving a prediction reflect real disease processes. A model can predict well without identifying any causal mechanism, and a biologically plausible explanation does not by itself prove causation; thus, each property needs its own evidence. Post hoc explanation methods, such as heatmaps laid over an image, can look reassuring while describing the model only loosely, and they may create false confidence at the level of the individual patient [9]. I agree with Rudin that, for high-stakes decisions, models should be interpretable from the outset rather than explained after the fact [10]. When an inherently interpretable model performs as well as a black box, it should be preferred. Where explanations are offered, they should be meaningful to the intended clinical user and decision, not only to data scientists, and should point to recognizable features such as a known pathway, a validated biomarker or a plausible radiological sign. Such plausibility guards against shortcut learning but is not proof of mechanism.

The second weakness concerns the data. The value of personalized medicine depends on the breadth of the population behind it. In 2016, approximately 81% of participants in genome-wide association studies were of European ancestry [11]; by 2021, this share had risen to 86%, whereas South Asian participants accounted for only 0.8% [12]. Polygenic risk scores built on these data are several times more accurate in people of European descent than in other groups [13]. It is worth distinguishing the bias already present in the health system that generates the data from the bias introduced by modeling choices. One widely used commercial algorithm in the United States relies on health care costs as a proxy for illness. Because unequal access meant that less money was spent

on black patients, they had to be considerably sicker than white patients were to receive the same risk score. This inequity originated in the health system, but the choice of label carried it into the algorithm; correcting it would have increased the percentage of black patients flagged for additional care from 17.7% to 46.5% [14]. Chest radiograph classifiers have similarly been shown to underdiagnose female, black, Hispanic and low-income patients, with compounded underdiagnosis in some intersectional subgroups, such as Hispanic female patients [15].

Representativeness is only part of the problem. Real-world performance also depends on the quality and consistency of the wider data ecosystem: whether records are complete and timely, whether diagnoses and laboratory tests are coded consistently, whether duplicate records and mismatched patient identifiers are resolved, and whether EHRs, laboratory information systems and diagnostic platforms can exchange data reliably. In practice, a model built on curated research data may encounter missing values, new assay platforms or local coding habits. Changes of this kind, known as dataset shifts, can degrade performance silently, without any change to the model itself [16]. Generalizability is therefore a property of the model together with the data system that feeds it.

These issues are sharper in low- and middle-income countries (LMICs), such as Pakistan. Their populations are poorly represented in the genomic and imaging datasets on which most commercial tools are trained [12], and a model validated in North America or Europe cannot be assumed to perform equally well in a South Asian hospital. Underrepresentation is compounded by fragmented digital systems, variable laboratory capacity, inconsistent data standards and few local datasets suitable for validation [17]. A scoping review revealed only 10 studies evaluating real-world applications of AI in health care in LMICs and reported problems with reliability, workflow fit and adaptation to the local context [18]. Rather than simply importing externally trained tools, these settings need locally governed data, regional validation networks and, where necessary, models developed or recalibrated for their own populations and health systems.

The third weakness is the absence of shared standards for evaluation. Clinical assessment should look beyond discrimination measures such as the AUROC. Calibration, the agreement between predicted and observed risk, is often unreported, yet a poorly calibrated model can mislead clinical decisions even when its discrimination is good [19]. Sensitivity, specificity and predictive values at the intended decision threshold, performance in relevant subgroups, and clinical utility (for example, net benefit estimated through decision curve analysis [20]) are equally important. The choice of metric should follow from the intended decision and from the relative harms of false-positive and false-negative results. The sepsis model described above illustrates this point: its low sensitivity and heavy alert burden are more important in practice than any single summary statistic is [8].

Validation should also be treated as continuing rather than as a single event. Prospective, multisite evaluation within real clinical workflows, ideally through randomized trials, remains essential, but it should include geographic, temporal and subgroup validation. Once deployed, models need continuous monitoring for drift in their inputs and performance. Clinical AI should follow a cycle of validation, deployment, monitoring, recalibration and revalidation rather than being regarded as permanently validated after one study [16,19]. Regulatory authorization is not a substitute for such evidence: of 521 AI-enabled medical devices authorized by the U.S. Food and Drug Administration, 226 (43%) lacked publicly reported clinical validation data [21]. The guidance largely exists. The Transparent Reporting of a Multivariate Prediction Model for Individual Prognosis Or Diagnosis + Artificial Intelligence (TRIPOD+AI) statement sets out how prediction models that use machine learning should be reported [22], the TRIPOD extension for

large language models (TRIPOD-LLM) applies the same principles to studies using these models [23], the Developmental and Exploratory Clinical Investigations of Decision support systems driven by Artificial Intelligence (DECIDE-AI) guideline covers the early clinical evaluation of decision support systems [24], and CONSORT-AI covers randomized trials [25]. What is missing is consistent enforcement. Journals can require adherence to these guidelines at submission, and regulators and hospital procurement committees can ask for external validation data, calibration and subgroup performance, and a post-deployment monitoring plan before a tool is adopted.

None of these steps are technically out of reach, but they require the research community to value careful evaluation as highly as it values headline performance. The ultimate aim of clinical AI is not better prediction but safer, more equitable and demonstrably better clinical decisions and patient outcomes. Tools that meet that standard will earn the trust of clinicians because they have been shown to work for all of their patients.

Abbreviations: AI, Artificial intelligence; AUROC, Area under the receiver operating characteristic curve; CONSORT-AI, Consolidated Standards of Reporting Trials–Artificial Intelligence; COVID-19, Coronavirus disease 2019; DECIDE-AI, Developmental and Exploratory Clinical Investigations of Decision support systems driven by Artificial Intelligence; EHR, Electronic health record; Low- and middle-income countries; TRIPOD+AI, Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis + Artificial Intelligence; TRIPOD-LLM, Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis–Large Language Models.

Author contributions: The author wrote and revised the manuscript, and approved the final version.

Use of generative AI and AI-assisted technologies: During the preparation of this work, the author used Claude Sonnet 4.5 (Anthropic) in order to improve the language and readability of the manuscript and to check that abbreviations were defined and used consistently. After using this tool, the author reviewed and edited the content as needed and takes full responsibility for the content of the published article.

Funding: This research received no specific grant from the public, commercial, or not-for-profit funding agencies.

Trial registration: Not applicable.

Ethics statement and consent to participate: Not applicable.

Consent to publication: Not applicable.

Data availability: Not applicable.

Acknowledgments: None.

Conflicts of interest: The author declares no conflicts of interest.

Publisher's note: Logixs Journals remains neutral concerning jurisdictional claims in its published subject matter, including maps and institutional affiliations.

References

- [1] Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. *Nat Med*. 2019;25:44-56. <https://doi.org/10.1038/s41591-018-0300-7>
- [2] Nagendran M, Chen Y, Lovejoy CA, Gordon AC, Komorowski M, Harvey H, et al. Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies. *BMJ*. 2020;368:m689. <https://doi.org/10.1136/bmj.m689>
- [3] Plana D, Shung DL, Grimshaw AA, Saraf A, Sung JJY, Kann BH. Randomized clinical trials of machine learning interventions in health care: a systematic review. *JAMA Netw Open*. 2022;5(9):e2233946. <https://doi.org/10.1001/jamanetworkopen.2022.33946>
- [4] Han R, Acosta JN, Shakeri Z, Ioannidis JPA, Topol EJ, Rajpurkar P. Randomised controlled trials evaluating artificial intelligence in clinical practice: a scoping review. *Lancet Digit Health*. 2024;6(5):e367-73. [https://doi.org/10.1016/S2589-7500\(24\)00047-5](https://doi.org/10.1016/S2589-7500(24)00047-5)

- [5] Kelly CJ, Karthikesalingam A, Suleyman M, Corrado G, King D. Key challenges for delivering clinical impact with artificial intelligence. *BMC Med.* 2019;17:195. <https://doi.org/10.1186/s12916-019-1426-2>
- [6] Zech JR, Badgeley MA, Liu M, Costa AB, Titano JJ, Oermann EK. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. *PLoS Med.* 2018;15(11):e1002683. <https://doi.org/10.1371/journal.pmed.1002683>
- [7] Wynants L, Van Calster B, Collins GS, Riley RD, Heinze G, Schuit E, et al. Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal. *BMJ.* 2020;369:m1328. <https://doi.org/10.1136/bmj.m1328>
- [8] Wong A, Otlés E, Donnelly JP, Krumm A, McCullough J, DeTroyer-Cooley O, et al. External validation of a widely implemented proprietary sepsis prediction model in hospitalized patients. *JAMA Intern Med.* 2021;181(8):1065-70. <https://doi.org/10.1001/jamainternmed.2021.2626>
- [9] Ghassemi M, Oakden-Rayner L, Beam AL. The false hope of current approaches to explainable artificial intelligence in health care. *Lancet Digit Health.* 2021;3(11):e745-50. [https://doi.org/10.1016/S2589-7500\(21\)00208-9](https://doi.org/10.1016/S2589-7500(21)00208-9)
- [10] Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat Mach Intell.* 2019;1:206-15. <https://doi.org/10.1038/s42256-019-0048-x>
- [11] Popejoy AB, Fullerton SM. Genomics is failing on diversity. *Nature.* 2016;538:161-4. <https://doi.org/10.1038/538161a>
- [12] Fatumo S, Chikowore T, Choudhury A, Ayub M, Martin AR, Kuchenbaecker K. A roadmap to increase diversity in genomic studies. *Nat Med.* 2022;28:243-50. <https://doi.org/10.1038/s41591-021-01672-4>
- [13] Martin AR, Kanai M, Kamatani Y, Okada Y, Neale BM, Daly MJ. Clinical use of current polygenic risk scores may exacerbate health disparities. *Nat Genet.* 2019;51:584-91. <https://doi.org/10.1038/s41588-019-0379-x>
- [14] Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science.* 2019;366(6464):447-53. <https://doi.org/10.1126/science.aax2342>
- [15] Seyyed-Kalantari L, Zhang H, McDermott MBA, Chen IY, Ghassemi M. Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. *Nat Med.* 2021;27:2176-82. <https://doi.org/10.1038/s41591-021-01595-0>
- [16] Finlayson SG, Subbaswamy A, Singh K, Bowers J, Kupke A, Zittrain J, et al. The clinician and dataset shift in artificial intelligence. *N Engl J Med.* 2021;385(3):283-6. <https://doi.org/10.1056/NEJMc2104626>
- [17] Kaushik A, Barcello C, Mandyam NK, Tan SY, Tromp J. Challenges and opportunities for data sharing related to artificial intelligence tools in health care in low- and middle-income countries: systematic review and case study from Thailand. *J Med Internet Res.* 2025;27:e58338. <https://doi.org/10.2196/58338>
- [18] Ciecierski-Holmes T, Singh R, Axt M, Brenner S, Barteit S. Artificial intelligence for strengthening healthcare systems in low- and middle-income countries: a systematic scoping review. *NPJ Digit Med.* 2022;5:162. <https://doi.org/10.1038/s41746-022-00700-y>
- [19] Van Calster B, McLernon DJ, van Smeden M, Wynants L, Steyerberg EW. Calibration: the Achilles heel of predictive analytics. *BMC Med.* 2019;17:230. <https://doi.org/10.1186/s12916-019-1466-7>
- [20] Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. *Med Decis Making.* 2006;26(6):565-74. <https://doi.org/10.1177/0272989X06295361>
- [21] Chouffani El Fassi S, Abdullah A, Fang Y, Natarajan S, Masroor AB, Kayali N, et al. Not all AI health tools with regulatory authorization are clinically validated. *Nat Med.* 2024;30:2718-20. <https://doi.org/10.1038/s41591-024-03203-3>
- [22] Collins GS, Moons KGM, Dhiman P, Riley RD, Beam AL, Van Calster B, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ.* 2024;385:e078378. <https://doi.org/10.1136/bmj-2023-078378>
- [23] Gallifant J, Afshar M, Ameen S, Aphinyanaphongs Y, Chen S, Cacciamani G, et al. The TRIPOD-LLM reporting guideline for studies using large language models. *Nat Med.* 2025;31:60-9. <https://doi.org/10.1038/s41591-024-03425-5>
- [24] Vasey B, Nagendran M, Campbell B, Clifton DA, Collins GS, Denaxas S, et al. Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. *Nat Med.* 2022;28:924-33. <https://doi.org/10.1038/s41591-022-01772-9>
- [25] Liu X, Cruz Rivera S, Moher D, Calvert MJ, Denniston AK, SPIRIT-AI and CONSORT-AI Working Group. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. *Nat Med.* 2020;26:1364-74. <https://doi.org/10.1038/s41591-020-1034-x>