American Journal of
Biomedical Science & Research

www.biomedgrid.com
ISSN: 2642-1747

Mini Review

Machine Learning in Precision Medicine: From Predictive Accuracy to Patient Benefit

Soumyadip Sarkar*

Independent Researcher, Kolkata, India

***Corresponding author:** Soumyadip Sarkar, Independent Researcher, Kolkata, India

To Cite This article: *Soumyadip Sarkar*, Machine Learning in Precision Medicine: From Predictive Accuracy to Patient Benefit. Am J Biomed Sci & Res. 2026 32(2) AJBSR.MS.ID.004137,* **DOI:** 10.34297/AJBSR.2026.32.004137

Received: September 04, 2026 **Published:** September 10, 2026

Abstract

Machine Learning (ML) can integrate imaging, molecular, physiological, and electronic health record data to support precision diagnosis and treatment. However, high retrospective discrimination is not evidence that an ML-enabled pathway benefits patients. This focused mini review examines the evidence gap and proposes an evidence-gated translation lifecycle. A review of 86 randomized trials found favourable primary endpoints in 81%, yet 54% evaluated diagnostic yield or performance, 63% were single center, and only 26% reported race or ethnicity. Prospective studies illustrate the gradient from performance to impact: autonomous diabetic-retinopathy screening achieved 87.2% sensitivity and 90.7% specificity without testing visual outcomes; the 105,934 participant MASAI mammography trial showed non-inferior interval-cancer detection; and a 74-unit pragmatic trial of ML deterioration surveillance reported lower in-hospital mortality but more unanticipated intensive-care transfers. External shift, poor calibration, shortcut learning, biased targets, and human-workflow effects can each invalidate apparent accuracy. Clinical ML should therefore be qualified through external silent evaluation, calibrated and subgroup-specific performance, decision-curve analysis, prospective impact trials, and versioned monitoring with rollback criteria. The clinically relevant intervention is the complete human–ML pathway, not the model alone.

Keywords: Machine learning, Precision medicine, Clinical prediction, Calibration, External validation, Randomized trial

Introduction

Precision medicine asks which action is appropriate for a particular patient, not merely which label is most probable. ML can learn useful representations from high-dimensional images, omics, waveforms, and longitudinal records, but its development objective is usually empirical prediction error. A model that minimizes $R(f) = n^{-1} \text{‘}[,i\ \ell(f(xi), yi)$ under historical data may rank risk well while producing unreliable probabilities or harmful decisions. Moreover, a prognostic model trained under prior treatment practice does not, by itself, estimate the counterfactual effect of choosing treatment A rather than B. Clinical translation must therefore evaluate the model, the action it triggers, and the surrounding human system (Figure 1).

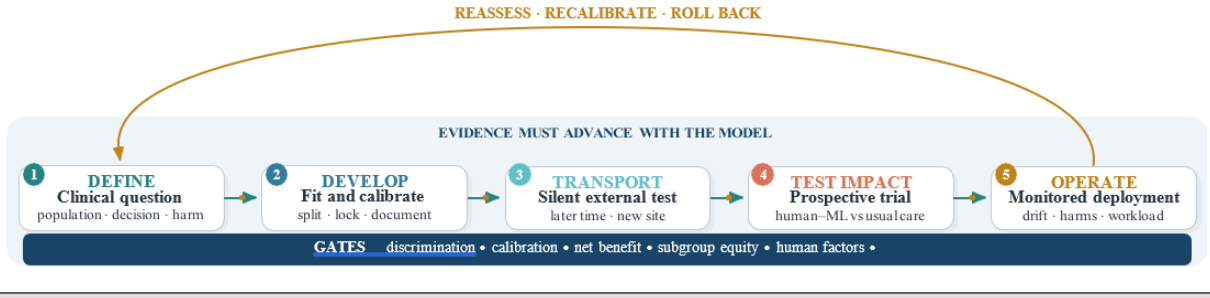

Figure 1: Evidence-gated clinical ML lifecycle. Progression is conditional; a failed gate requires redesign or withdrawal. Monitoring closes the validation loop.

Scope and Search Approach

We conducted a focused narrative search of PubMed and multidisciplinary scholarly indexes for English-language peer-reviewed studies combining "machine learning" with clinical prediction, calibration, external validation, randomized evaluation, fairness, and monitoring. We prioritized primary human studies, randomized trials, reviews of trials, and reporting or deployment guidance, with backward and forward citation checking. This was not a registered systematic review, and no meta-analysis was performed.

The Clinical Evidence Gradient

Han, et al., identified 86 randomized trials published through 14 November 2023. Although 70 (81%) reported a favorable primary endpoint, 46 (54%) focused on diagnostic yield or performance, 54 (63%) were single center, and race or ethnicity was reported in only 22 (26%) [1]. Publication bias was also plausible. These findings warn against treating the number of positive trials as the number demonstrating broad patient benefit (Figure 2).

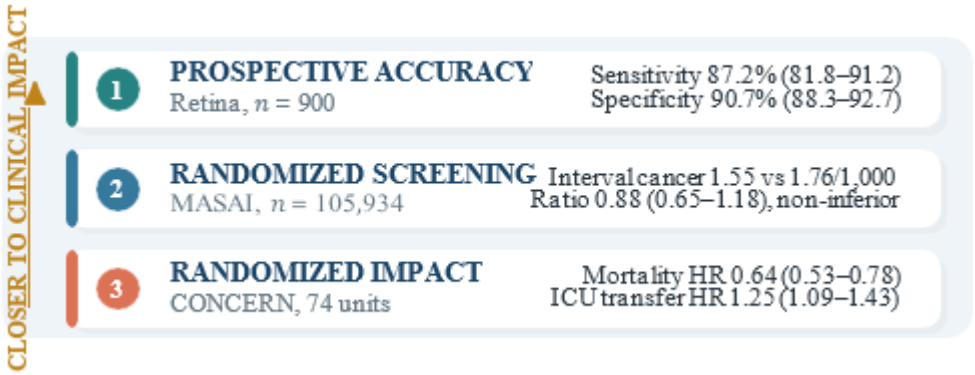

Figure 2: Three evidence levels in clinical ML. Values are 95% confidence intervals from prospective retina screening, MASAI, and CONCERN [2- 4].

Performance remains an intermediate endpoint: retina screening did not test visual outcomes, and MASAI did not establish mortality benefit. CONCERN linked a complete ML surveillance pathway to patient outcomes, but its two-system, COVID era setting limits transport. These results support specific interventions, not an interchangeable class effect of ML.

Why High Accuracy Can Fail

Discrimination ranks cases; calibration asks whether estimated risk matches observed frequency, ideally Pr (Y=1 |p= p) ≈ p [5-7]. Neither measures the value of acting at a threshold. Figure 3 separates three failures that a headline AUC can conceal (Figure 3).

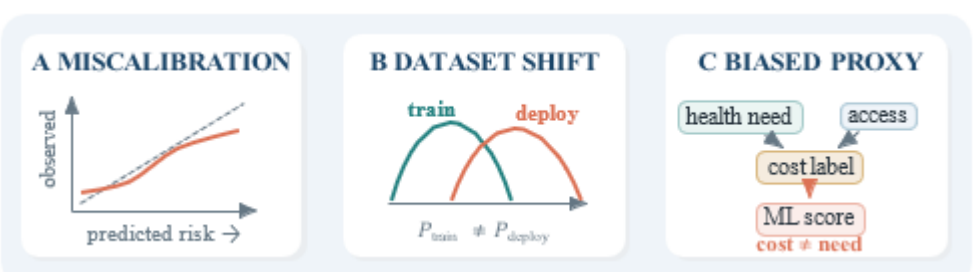

Figure 3: Why accuracy can fail: unreliable probabilities, shifted inputs, and a target that encodes access rather than need [5-7]. Original schematic.

Transport can fail when prevalence, acquisition, coding, or workflow changes. In 158,323 radiographs, a pneumonia detector fell from internal AUC 0.931 (95% CI 0.927–0.936) to external AUC 0.815 (0.745–0.885; p=0.001), while hospital source was identified with 99.95–99.98% accuracy [5]. Target choice can also encode inequity: replacing cost with health need in a population-management algorithm would have raised Black patients selected for additional care from 17.7% to 46.5% [6]. Fairness therefore begins with the clinical question, label, missingness, and allocation process.

An Evidence-Gated ML Lifecycle

First, prespecify the intended population, decision, out-come, time horizon, action threshold, and acceptable harms. Partition patients rather than records; freeze preprocessing, features, parameters, and thresholds before evaluation; and document missing-data handling and leakage controls. Report discrimination, calibration-in-the-large, calibration slope, threshold-specific errors, net benefit, confidence intervals, and clinically justified subgroup results. TRIPOD+AI provides a reporting structure for development and external evaluation [8].

Second, test the locked pathway silently across later time periods and genuinely external sites. Evaluate both performance and resource consequences before recommendations reach clinicians. Early live studies should characterize usability, automation bias, overrides, alert fatigue, and work-flow adaptation; DECIDE-AI provides relevant guidance [9]. When predictions change care and material benefit or harm is plausible, randomize the complete intervention against current practice using patient centered endpoints. Third, treat deployment as continued evaluation. Version the model and data pipeline; monitor input shift, calibration, net benefit, subgroup error, latency, workload, overrides, and adverse events; define triggers for investigation, recalibration, rollback, or retirement. FUTURE-AI frames these obligations around fairness, universality, traceability, usability, robustness, and explainability [10]. Governance must assign owners and preserve clinician patient authority.

Conclusion

ML is useful in selected workflows, and randomized evidence shows that some ML-enabled pathways can improve out-comes. Benefit cannot be inferred from AUC, internal validation, or novelty. A credible claim requires calibrated external performance, favourable decision consequences, prospective comparison, equitable operation, and monitoring. The intervention to validate is the human–ML pathway, not the algorithm alone.

Declarations

Acknowledgments and funding

None.

Conflict of interest

The author(s) declare no conflict of interest.

Ethics and data availability

No participants were enrolled and no new data were generated; ethics approval and consent were not applicable.

References

1. Han R, Acosta JN, Shakeri Z, Ioannidis JPA, Topol EJ, et al. (2024) Randomised controlled trials evaluating artificial intelligence in clinical practice: a scoping review. Lancet Digit Health 6(5): e367–e373.
2. Abràmoff MD, Lavin PT, Birch M, Shah N, Folk JC (2018) Pivotal trial of an autonomous AI-based diagnostic system for detection of diabetic retinopathy in primary care offices. NPJ Digit Med 1(1): 39.
3. Gommers JJM, Hernström V, Josefsson V, Sartor H, et al. (2026) Interval cancer, sensitivity, and specificity comparing AI-supported mammography screening with standard double reading without AI in the MASAI study. Lancet 407(10527): 505–514.
4. Rossetti SC, Dykes PC, Knaplund C, Cho S, Withall J, et al. (2025) Real-time surveillance system for patient deterioration: a pragmatic cluster-randomized controlled trial. Nat Med 31(6):1895–1902.
5. Zech JR, Badgeley MA, Liu M, Costa AB, Titano JJ, et al. (2018) Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. Multicenter Study 15(11): e1002683.
6. Obermeyer Z, Powers B, Vogeli C, Mullainathan S (2019) Dissecting racial bias in an algorithm used to manage the health of populations. Science 366(6464):447–453.
7. Van Calster B, McLernon DJ, van Smeden M, Wynants L, Steyerberg EW (2019) Calibration: the Achilles heel of predictive analytics. BMC Med 17(1): 230.
8. Collins GS, Moons KGM, Dhiman P, Riley RD, Beam AL, Van Calster B, et al. (2024) TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ 385: e078378.
9. Vasey B, Nagendran M, Campbell B, Clifton DA, Collins GS, et al. (2022) Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. Practice Guideline 28(5): 924–933.
10. Lekadir K, Frangi AF, Porras AR, Glocker B, Cintas C, et al. (2025) FUTURE-AI: international consensus guideline for trustworthy and deployable artificial intelligence in healthcare. BMJ 388: e081554.