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

***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 does not establish that an ML-enabled clinical pathway benefits patients. This focused mini review examines that evidence gap and proposes an evidence-gated translation lifecycle. A scoping review identified 86 randomized trials, of which 70 (81%) had a favorable primary-endpoint result; 46 (54%) used diagnostic yield or performance as the primary endpoint, 54 (63%) were single-center, and only 22 (26%) reported race or ethnicity. Prospective studies illustrate the gradient from predictive performance to clinical impact: autonomous diabetic-retinopathy assessment achieved 87.2% sensitivity and 90.7% specificity among 819 analyzable participants without testing visual outcomes; MASAI randomized 105,934 women and found a non-inferior interval-cancer rate with AI-supported mammography screening; and a 74-unit pragmatic cluster-randomized trial of ML deterioration surveillance found lower in-hospital mortality but a higher hazard of unanticipated intensive-care transfer. Dataset shift, miscalibration, confounding or shortcut signals, biased targets, and human-workflow effects can undermine transportability or clinical benefit despite high apparent accuracy. A defensible translation pathway therefore requires evidence beyond discrimination, including external and subgroup-specific validation, calibration, threshold-based clinical utility, prospective impact evaluation when appropriate, and versioned post-deployment monitoring. 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 (Figure 1).

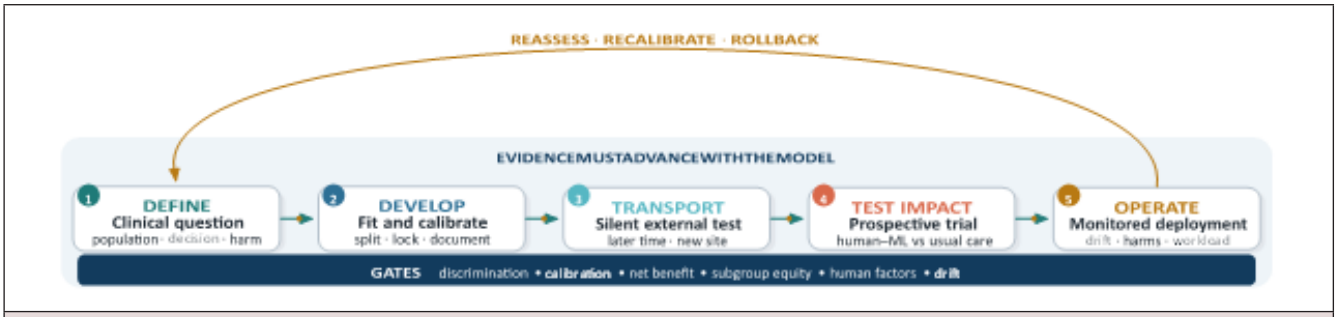

Figure 1: Evidence-gated clinical ML lifecycle. Progression is conditional; a failed gate requires redesign or withdrawal. Monitoring closes the validation loop.

Introduction

Precision medicine ultimately concerns choosing actions for particular patients, not merely assigning labels. ML can learn useful representations from high-dimensional images, omics, waveforms, and longitudinal records, but model development commonly minimizes predictive loss. For observations $(x_i, y_i)i^n_{=1}$, empirical risk can be written as $R_n(f)=1/n\sum i=1^n \ell(f(x_i), y_i)$. Low empirical or held-out prediction error does not by itself guarantee calibrated probabilities or beneficial downstream decisions [1,2]. Likewise, a prognostic model trained under historical treatment practice does not, by itself, identify the causal contrast between treatment A and treatment B for an individual patient [3]. Clinical translation must therefore evaluate the model, the action it triggers, and the surrounding human system.

Scope and Search Approach

We conducted a focused narrative search of PubMed and multidisciplinary scholarly indexes, supplemented by backward and forward citation checking. Search concepts combined "machine learning" or "artificial intelligence" with clinical prediction, calibration, external validation, randomized evaluation, fairness, monitoring, and related clinical-translation terms. We prioritized peer-reviewed primary human studies, randomized trials, reviews of trials, and reporting or deployment guidance. This was not a registered systematic review, no formal risk-of-bias assessment was performed for this mini review, and no meta-analysis was undertaken.

The Clinical Evidence Gradient

Han, et al., identified 86 eligible randomized trials in a scoping-review search conducted. Seventy (81%) had a favorable primary-endpoint result, including statistically significant improvement or established non-inferiority; 46 (54%) used diagnostic yield or performance as the primary endpoint, 54 (63%) were single-center, and race or ethnicity was reported in only 22 (26%) [4]. The authors also noted likely publication bias. These findings warn against equating the number of favorable trials with the number demonstrating broad patient benefit (Figure 2).

Figure 2: Three evidence levels in clinical ML. Retina estimates are from the 819 analyzable participants (900 enrolled); MASAI randomized 105,934 women; CONCERN randomized 74 clinical units. Values in parentheses are 95% confidence intervals [5,6,7].

Performance remains an intermediate endpoint: the retina study evaluated diagnostic accuracy rather than visual outcomes, and MASAI established non-inferiority for interval cancer rate rather than mortality benefit. CONCERN linked a complete ML surveillance pathway to patient outcomes, but evaluation across only two health systems constrains transportability. A 2026 author correction revised length-of stay and 30-day readmission analyses but did not alter the mortality or unanticipated ICU-transfer hazard ratios shown in Figure 2 [8]. These results support specific interventions, not an interchangeable class effect of ML. Why High Accuracy Can Fail Discrimination measures how well predictions rank individuals with different outcomes; calibration concerns agreement between predicted probabilities and observed event frequencies. Ideal calibration satisfies Pr(Y = 1 | p = p) = p b [1]. Neither discrimination nor calibration alone quantifies the clinical value of acting at a particular risk threshold; decision-curve analysis addresses that question through threshold-weighted net benefit [2]. 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 [1,9,10].

Transport can fail when prevalence, acquisition, coding, or workflow changes. In 158,323 chest radiographs, a pneumonia CNN trained and tested on pooled MSH and NIH data achieved internal AUC 0.931 (95% CI 0.927–0.936) but external AUC 0.815 (0.745–0.885; $p = 0.001$) at IU. A separate site-classification CNN identified NIH, MSH, and IU origin with 99.95%, 99.98%, and 95.59% accuracy, respectively [9]. Target choice can also encode inequity: in the population-management algorithm studied by *Obermeyer, et al.,* health-care cost was used as a proxy for health need; the authors estimated that remedying the resulting disparity would increase the proportion of Black patients selected for additional care from 17.7% to 46.5% [10]. Fairness therefore begins with the clinical question, label, missingness, and allocation process.

An Evidence-Gated ML Lifecycle

First, prespecify the intended population, decision, outcome, time horizon, action threshold, and acceptable harms. When patients contribute multiple records, keep records from the same patient within the same data partition; use temporal, site-level, or other separation when required by the intended deployment setting. Lock preprocessing, features, model parameters, and decision thresholds before a definitive evaluation, and document missing-data handling and leakage controls. Report discrimination, calibration-in-the-large, calibration slope, threshold-specific errors, confidence intervals, and clinically justified subgroup results; when a prediction will trigger an action, evaluate clinical utility across relevant thresholds using net benefit or another prespecified decision analytic measure [2]. TRIPOD+AI provides a reporting structure for development and external evaluation [11].

Second, evaluate the locked pathway prospectively across later time periods and genuinely external sites, using silent deployment when appropriate to characterize transportability before model output sinfluencecare. Earlylivestudiesshould characterize usability, automation bias, overrides, alert fatigue, workflow adaptation, and resource consequences; DECIDE-AI provides relevant guidance [12]. When model outputs materially change care, randomized evaluation of the complete human–ML intervention against current practice, when feasible and ethically appropriate, can provide direct evidence about causal effects on patient-centered outcomes. Third, treat deployment as continued evaluation. Version the model and data pipeline; monitor input shift, calibration, clinical utility, subgroup error, latency, workload, overrides, and adverse events; and prespecify triggers for investigation, recalibration, rollback, or retirement. FUTURE-AI frames lifecycle governance around fairness, universality, traceability, usability, robustness, and explainability, including external evaluation, local validity, auditing, logging, and updating [13]. Governance should assign accountable owners while preserving appropriate clinician oversight and patient autonomy.

Conclusion

ML is useful in selected workflows, and randomized evidence shows that some ML-enabled pathways can improve patient or care-process outcomes. Clinical benefit cannot be inferred from AUC, internal validation, or model novelty alone. A credible claim of patient benefit should be supported by calibrated external performance, favorable decision consequences, prospective comparison when appropriate, equitable operation, and continued monitoring. When model outputs alter care, the intervention to evaluate is the human–ML pathway, not the algorithm in isolation.

Declarations

Acknowledgments and Funding

None.

Conflict of Interest

The author declares no conflict of interest.

Ethics and Data Availability

No participants were enrolled and no new data were generated; ethics approval and informed consent were therefore 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.