Available Online: 15 July 2026 DOI: 10.54254/3049-5458/2026.35444

Applications and limitations of machine learning in clinical biostatistics: a narrative review

Xiuyi Wei

School of Global Public Health, New York University, New York, USA

xw4000@nyu.edu

Abstract. Clinical biostatistics has traditionally relied on regression-based models to explain associations, estimate risks, and support medical decision-making. The growth of electronic health records, imaging data, omics data, and follow-up data has made clinical information larger, less tidy, and more difficult to model with only classical methods. Machine learning is increasingly used in this setting because it can capture non-linear patterns and handle high-dimensional predictors. This narrative review summarizes applications of supervised learning, unsupervised learning, and deep learning in clinical diagnosis, prognosis prediction, patient stratification, and biomarker discovery. Literature was selected from PubMed, Web of Science, and Google Scholar, with emphasis on studies and reporting guidelines published from 2019 to 2025. This review finds that machine learning is useful when the clinical question is clear, data quality is acceptable, and validation is strict. However, a stronger algorithm does not automatically become a better clinical tool. Main limitations include weak interpretability, biased training data, poor transportability, overreliance on Area Under the Curve (AUC), and incomplete reporting. Machine learning should therefore be treated as a complement to traditional biostatistics rather than a simple replacement.

Keywords: machine learning, clinical biostatistics, prediction model, external validation, medical AI

1. Introduction

Clinical biostatistics provides the basic language for medical evidence. Logistic regression, cox regression, generalized linear models, and survival analysis remain central tools for estimating associations, building risk scores, and explaining disease outcomes. In recent years, clinical data have become more complex because hospitals now store large amounts of electronic health records, imaging files, waveforms, notes, and omics measurements. Machine learning has entered this field because it can model non-linear relations and interactions that are difficult to pre-specify in traditional models [1]. Artificial intelligence is also viewed as part of a wider movement toward high-performance medicine, where algorithms may help with diagnosis, triage, monitoring, and treatment selection [2].

The important issue is not whether an algorithm looks advanced. The practical issue is whether it improves clinical decisions beyond simpler and more transparent methods. A systematic review found no clear overall performance advantage of machine learning over logistic regression in many clinical prediction studies [3]. This finding does not make machine learning useless, but it makes the hype look too easy. A model with a

higher AUC in one dataset may still be poorly calibrated, unfair across subgroups, or impossible to use in a busy clinic. This review mainly focuses on applications, evaluation standards, and limitations of machine learning in clinical biostatistics. The aim is to clarify where machine learning adds value, where traditional biostatistics remains stronger, and what conditions are needed before a model can be trusted in practice.

2. Methods

2.1. Literature search and selection

This article uses a narrative literature review design. The purpose is to synthesize major themes rather than pool effect sizes. PubMed, Web of Science, and Google Scholar were searched for English-language literature published mainly from 2019 to 2025. Search terms included machine learning, clinical prediction model, external validation, medical artificial intelligence, clinical biostatistics, biomarker discovery, interpretability healthcare, TRIPOD-AI, PROBAST-AI, CONSORT-AI, and SPIRIT-AI. Studies were included when they discussed clinical diagnosis, prognosis prediction, patient stratification, biomarker discovery, reporting standards, validation, or bias in medical AI. Articles were excluded when they only explained algorithms without a clinical setting, were conference abstracts, were duplicates, or lacked accessible full text.

2.2. Synthesis approach

The selected literature was grouped into four themes. The first theme was clinical application, including diagnosis, prognosis, and biomarker work. The second theme was model type, including supervised learning, unsupervised learning, and deep learning. The third theme was model evaluation, including discrimination, calibration, external validation, and decision-curve analysis. The fourth theme was translation, including interpretability, fairness, reporting, and implementation. Interpretability literature was considered because clinical users need more than a final risk number [4]. Biomarker studies were reviewed because omics data are a typical high-dimensional setting where machine learning is attractive [5]. Recent reporting and risk-of-bias tools were also included because incomplete reporting can make a published model hard to judge or reproduce [6-8]. Literature on bias, clinical trials, and early-stage Artificial Intelligence (AI) evaluation was used to connect model development with real clinical deployment [9-12].

3. Applications of machine learning in clinical biostatistics

3.1. Supervised learning for diagnosis and prognosis

Supervised learning is the most common machine learning approach in clinical biostatistics. It uses labeled outcomes to train models for classification or prediction. In diagnosis, it may combine symptoms, laboratory values, imaging features, and medical history to estimate disease probability. In prognosis, it may predict mortality, readmission, complications, or disease progression. Common tools include logistic regression used as a baseline, random forests, Support Vector Machines (SVM), gradient boosting, and neural networks.

The main advantage is flexibility. Supervised learning can handle many predictors and can capture interactions that researchers may not define in advance. This is useful when the outcome is influenced by several weak signals rather than one dominant risk factor. The weakness is also familiar. A model can fit local patterns in the development dataset and then lose performance in another hospital. For clinical prediction, internal accuracy is not enough. The model must keep acceptable calibration and discrimination after external validation, and its output must match a real decision point.

3.2. Unsupervised learning for patient stratification

Unsupervised learning does not begin with a known outcome label. It searches for hidden structures inside the data. In clinical research, it is often used to identify disease subtypes, patient clusters, or biomarker patterns. Methods include clustering, Principal Component Analysis (PCA), t-distributed Stochastic Neighbor Embedding (t-SNE), and Uniform Manifold Approximation and Projection **(**UMAP). In omics research, these methods can reduce thousands of variables into smaller patterns that may suggest biological mechanisms or patient groups.

The limitation is that a cluster is not automatically meaningful. A computer can divide patients into groups, but the groups may not be stable, clinically interpretable, or useful for treatment. Good unsupervised research therefore needs follow-up analysis. Clusters should be tested in independent data, compared with known clinical features, and examined for outcome differences. Without these steps, patient stratification may become an attractive picture rather than usable evidence.

3.3. Deep learning and multimodal clinical data

Deep learning is most visible in imaging, pathology, electrocardiography, and other high-dimensional data. Convolutional Neural Networks (CNN) can detect visual patterns in Computed Tomography (CT), Magnetic Resonance Imaging (MRI), retinal images and histology slides. Recurrent Neural Networks (RNN) and transformer-based models can process time-series data, clinical notes, and multimodal records. The attraction is feature learning. Traditional modeling often requires researchers to define variables before analysis, while deep learning can learn representations directly from raw or semi-processed data.

The cost is high. Deep learning usually requires large datasets, careful labeling, computing resources and strict monitoring. It is also harder to explain. This does not mean that deep learning should be avoided, but it does mean that stronger governance is needed. In a clinical setting, a wrong prediction is not only a technical error. It can change workload, patient trust, and treatment decisions. As summarized in Table 1, different machine learning approaches contribute to different clinical tasks, and the choice of method should depend on the research question, data structure, and clinical objective.

Table 1. Main machine learning approaches and clinical uses

Approach	Common tools	Clinical uses	Main concern
Supervised learning	Logistic regression, random forest, SVM, XGBoost, neural networks	Diagnosis, prognosis, readmission, mortality prediction	Needs calibration and external validation
Unsupervised learning	Clustering, PCA, t-SNE, UMAP	Patient stratification, subtype discovery, omics patterns	Clusters must be stable and clinically interpretable
Deep learning	CNN, RNN, transformer models	Imaging, pathology, text, time-series and multimodal data	Requires large data, governance, and explainability

4. Evaluation, limitations and translation

4.1. Prediction performance beyond AUC

Many clinical machine learning papers emphasize AUC. AUC measures discrimination, or how well a model ranks patients with and without an outcome. It is useful, but it is not enough. A model can have a high AUC and still give risk estimates that are too high or too low. Calibration is therefore essential because many clinical decisions depend on absolute risk. For example, a clinician needs to know whether a patient is near a treatment threshold, not only whether the patient is ranked above another patient.

External validation is also necessary. Cross-validation and bootstrapping test performance inside the available dataset, but they cannot fully show whether the model works in another hospital, another country, or another time period. Decision-curve analysis is also helpful because it connects prediction with clinical consequences. A model should improve decisions compared with current practice, not just produce better statistics on paper.

4.2. Interpretability and clinical reasoning

Interpretability is linked to trust, safety, and accountability. Clinicians do not need every mathematical detail, but they need enough information to know when a model is reasonable and when it may be misleading. Feature importance, partial dependence plots, SHapley Additive exPlanations (SHAP) values, and simpler surrogate models can help, although they also have limits.

Explanation should not be confused with causation. A model may learn that a medication predicts poor prognosis because sicker patients receive that medication, not because the medication is harmful. Traditional biostatistics is often stronger when the question is causal or mechanistic. Machine learning is often stronger when the goal is pure prediction. The method should follow the question.

4.3. Bias, generalizability and fairness

Bias can enter at many stages. Training data may underrepresent some groups, labels may reflect unequal access to care, and missing data may be related to insurance status, language, or hospital workflow. If these problems are ignored, a model may reproduce or amplify existing inequalities. Subgroup analysis is therefore not optional. Performance should be checked across clinically relevant groups, including age, sex, race, ethnicity, socioeconomic position, and care setting when such variables are available and ethically appropriate.

Table 2. Practical requirements for clinical machine learning models

Requirement	Purpose	Practical meaning
External validation	Test transportability	Evaluate the model in data from another setting or period
Calibration	Check absolute risk accuracy	Compare predicted risks with observed outcomes
Subgroup evaluation	Detect unfair performance	Report performance across clinically important groups
Transparent reporting	Support reproducibility	Describe data, predictors, missing data, model building, and validation
Implementation monitoring	Maintain safety	Track alert burden, drift, patient impact, and clinician response

Generalizability is closely related. A model trained in a large academic center may not work in a community hospital. A model trained before a change in clinical practice may fail after that change. This means model monitoring is needed after deployment. Clinical models should be treated as living tools that require updating, not as finished products after publication. Table 2 highlights that model evaluation in clinical machine learning extends beyond predictive accuracy and requires attention to validation, calibration, interpretability, and implementation.

4.4. Reporting and implementation standards

Reporting guidelines matter because incomplete reporting makes a model hard to evaluate. A paper should describe the data source, eligibility criteria, missing data handling, predictors, outcome definition, modeling process, validation strategy, and calibration results. TRIPOD+AI and PROBAST+AI extend earlier prediction-model standards for artificial intelligence methods. CONSORT-AI and SPIRIT-AI add details for trials of AI interventions, and DECIDE-AI supports early-stage clinical evaluation.

Implementation is a separate challenge. A statistically strong model can still fail if it causes alert fatigue, slows clinicians down, gives unclear responsibility, or lacks integration with electronic health records. Real clinical value depends on workflow fit, human oversight, and patient safety monitoring.

4.5. Relationship with traditional biostatistics

Machine learning and traditional biostatistics are sometimes presented as competitors, but that framing is not very useful. Traditional methods are transparent, efficient for smaller datasets, and strong for effect estimation. Machine learning is useful for complex prediction, high-dimensional data, and pattern recognition. The better approach is to choose the method according to the research question. If the goal is to estimate the effect of an exposure, a traditional model with careful design may be stronger. If the goal is to predict risk from many weak and non-linear signals, machine learning may be stronger.

Clinical researchers also need to understand both sides. A person who only knows algorithms may miss bias, confounding, and validation problems. A person who only knows classical models may miss useful prediction tools. Future clinical biostatistics should combine statistical reasoning with machine learning skills.

5. Conclusion

Machine learning has become an important part of clinical biostatistics because clinical data are no longer limited to a few clean variables in a small table. Electronic health records, imaging data, omics data, and repeated measurements create prediction problems that are difficult for traditional methods alone. Supervised learning can support diagnosis and prognosis prediction, unsupervised learning can help identify patient subgroups, and deep learning can process images, text, and multimodal data. These contributions are real, especially when the clinical question is practical and the dataset is rich enough.

At the same time, machine learning does not remove the basic problems of clinical research. A model is only as good as its data, outcome definition, validation process, and clinical use case. High AUC does not prove safety. A complex algorithm does not prove clinical value. A published model does not prove that implementation will work. The main weaknesses are interpretability, bias, poor transportability, limited external validation, and weak reporting. These problems are not small technical details. They decide whether a model helps patients or simply becomes another paper.

The most reasonable position is that machine learning should complement traditional biostatistics. Traditional methods remain essential for causal thinking, transparent modeling, and study design. Machine

learning adds value when the goal is prediction across complex data structures. Future work should focus less on presenting a new algorithm and more on making models clinically usable. Better data quality, transparent reporting, external validation, subgroup fairness checks, calibration assessment, decision-curve analysis, and post-deployment monitoring should become routine. In that direction, clinical machine learning can move from impressive development results toward tools that are testable, safe, and useful in real medical settings.

References

[1] Rajkomar, A., Dean, J., & Kohane, I. S. (2019). Machine learning in medicine. *New England Journal of Medicine, 380*(14), 1347–1358.

[2] Topol, E. J. (2019). High-performance medicine: The convergence of human and artificial intelligence. *Nature Medicine, 25*(1), 44–56.

[3] Christodoulou, E., Ma, J., Collins, G. S., Steyerberg, E. W., Verbakel, J. Y., & Van Calster, B. (2019). A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *Journal of Clinical Epidemiology, 110*, 12–22.

[4] Stiglic, G., Kocbek, P., Fijacko, N., Zitnik, M., Verbert, K., & Cilar, L. (2020). Interpretability of machine learning-based prediction models in healthcare. *WIREs Data Mining and Knowledge Discovery, 10*(5), e1379.

[5] Glaab, E. (2021). Biomarker discovery studies for patient stratification using machine learning analysis of omics data. *BMJ Open, 11*(12), e053674.

[6] Collins, G. S., Moons, K. G. M., Dhiman, P., Riley, R. D., Beam, A. L., Van Calster, B., … Reitsma, J. B. (2024). TRIPOD+AI statement: Updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ, 385*, e078378.

[7] Wolff, R. F., Moons, K. G. M., Riley, R. D., Whiting, P. F., Westwood, M., Collins, G. S., … PROBAST Group. (2019). PROBAST: A tool to assess the risk of bias and applicability of prediction model studies. *Annals of Internal Medicine, 170*(1), 51–58.

[8] Moons, K. G. M., Damen, J. A. A. G., Kaul, T., Dhiman, P., Hooft, L., Reitsma, J. B., … Collins, G. S. (2025). PROBAST+AI: A tool to assess risk of bias and applicability of prediction model studies using artificial intelligence. *BMJ, 388*, e082505.

[9] Cross, J. L., Choma, M. A., & Onofrey, J. A. (2024). Bias in medical AI: Implications for clinical decision-making. *PLOS Digital Health, 3*(11), e0000651.

[10] Liu, X., Rivera, S. C., Moher, D., Calvert, M. J., & Denniston, A. K. (2020). Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: The CONSORT-AI extension. *Nature Medicine, 26*(9), 1364–1374.

[11] Rivera, S. C., Liu, X., Chan, A. W., Denniston, A. K., & Calvert, M. J. (2020). Guidelines for clinical trial protocols for interventions involving artificial intelligence: The SPIRIT-AI extension. *The Lancet Digital Health, 2*(10), e549–e560.

[12] Vasey, B., Nagendran, M., Campbell, B., Clifton, D. A., Collins, G. S., Denaxas, S., … McCulloch, P. (2022). Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. *Nature Medicine, 28*(5), 924–933.