



ISSN : 2347 - 2243

*Indo - American Journal of
Life Sciences and Biotechnology*



www.iajlb.com

Email : editor@iajlb.com or iajlb.editor@gmail.com



Machine Learning–Based Prediction Models for Early Detection and Risk Stratification of Cardiovascular Diseases: A Systematic Review

Dr. Kapil Khanna¹, Dr. Harsh Rameshbhai Goti², Dr. Tarunkumar Gulabsinh Baria^{3*},
Dr. Sameer Srivastava⁴, Dr. Anupam Tyagi⁵

¹MD, PGDCCP, Santosh Deemed to be University, Ghaziabad, Uttar Pradesh, India.

²Junior Doctor, Department of Neurology, Dave Neurology Clinic, Surat, Gujarat, India.

^{3*}Junior Resident, Department of General Surgery, GMERS Medical College, Panchmahal, Godhra, Gujarat, India.

⁴Professor, Department of Physiology, Maharishi Vashishtha Autonomous State Medical College, Basti, Uttar Pradesh, India.

⁵Associate Professor, Department of Pharmacology, Maharishi Vashishtha Autonomous State Medical College, Basti, Uttar Pradesh, India.

Corresponding Author: Dr. Tarunkumar Gulabsinh Baria

Junior Resident, Department of General Surgery, GMERS Medical College, Panchmahal, Godhra, Gujarat, India.

Email: samirbaria505@gmail.com

Received: 02.09.26, Accepted: 08.10.26

ABSTRACT

Background: Machine-learning (ML) models can use high-dimensional electronic health record, imaging, waveform and biomarker data to estimate cardiovascular risk, but apparent gains in discrimination may be offset by poor calibration, overfitting and limited external validation.

Objectives: To compare ML-based cardiovascular risk prediction with conventional risk scores, evaluate model quality and external validity, and quantify discrimination in studies reporting directly comparable AUC estimates.

Methods: A PRISMA-structured search framework covered PubMed, Scopus, Embase and Cochrane through September 2026. Cohort studies developing or validating ML models for incident cardiovascular disease or ASCVD risk were eligible when they reported clinically interpretable discrimination and a conventional comparator. PROBAST was used as the primary quality framework. Random-effects models separately pooled reported AUCs for ML models and conventional scores in three large comparative cohorts.

Results: The three directly comparable cohorts included hundreds of thousands of participants. Pooled ML AUC was 0.791 (95% CI 0.758–0.823; I²=98.6%), whereas pooled conventional-score AUC was 0.738 (95% CI 0.723–0.752; I²=92.2%). The descriptive difference was about 0.053, but paired inference was not performed because within-study covariance was unavailable. Larger systematic reviews confirm modest average discrimination gains but identify high risk of bias, inconsistent calibration reporting and scarce independent external validation.

Conclusion: ML can improve discrimination and broaden risk estimation beyond conventional variable sets, but current evidence does not support uncritical replacement of established



scores. Independent external validation, calibration, decision-curve analysis, fairness assessment and prospective impact studies are required.

Keywords: Machine Learning, Cardiovascular Disease, Risk Prediction, Risk Stratification, Probst, Electronic Health Records, Meta-Analysis.

INTRODUCTION

Cardiovascular disease remains a leading cause of morbidity and mortality, and prevention depends on identifying individuals whose absolute risk is high enough to justify intensive risk-factor modification. Conventional scores such as the pooled cohort equations, Framingham-derived models and QRISK were designed for interpretability and clinical implementation using a limited set of predictors. Their limitations include transportability across populations, missing or out-of-range variables, non-linear relationships and imperfect calibration in changing populations [1-2].

Machine learning offers several theoretical advantages. Penalized regression, random forests, gradient boosting, neural networks and survival-learning methods can model complex interactions, accommodate larger predictor sets and extract information from longitudinal electronic health records. Weng and colleagues showed that several ML approaches improved 10-year cardiovascular risk discrimination compared with an ACC/AHA algorithm in 378,256 primary-care patients [3]. Alaa and colleagues subsequently used automated ML in 423,604 UK Biobank participants and reported higher discrimination than Framingham-based prediction [4].

However, better apparent AUC does not necessarily mean better clinical decisions. Complex models can overfit to local coding patterns, use variables that are proxies for healthcare access, and become miscalibrated when applied to a new setting. A 2024 systematic review identified 486 AI cardiovascular prediction models across 79 articles but found no

model with independent external validation in its evidence set and judged all models at high risk of bias using PROBAST [5].

External validation is particularly important because risk prediction is a calibration-sensitive task. A model may rank patients correctly yet systematically overestimate or underestimate absolute risk, leading to inappropriate treatment thresholds. In addition, improvements in AUC can be statistically significant but clinically modest. Net benefit, reclassification, calibration-in-the-large, calibration slope and performance across demographic groups are therefore necessary complements to discrimination.

The present systematic review and meta-analysis focuses on ML for early detection and risk stratification of cardiovascular disease, with particular attention to direct comparisons against conventional scores, model validation, bias and implementation readiness. The objective is not to identify a single “best” algorithm but to determine the strength and limitations of the evidence supporting ML-based risk prediction.

PICO/Peco Framework

Because most prediction-model studies are observational, a PECO/PICO hybrid was used: **Population:** adults without established cardiovascular disease at baseline or populations undergoing cardiovascular risk assessment; **Exposure/Intervention:** an ML-based prediction model; **Comparator:** conventional risk score, regression model or clinical benchmark; **Outcomes:** AUC/C-statistic, calibration, sensitivity/specificity, reclassification, decision-curve net benefit and external validation. Studies predicting



short-term acute outcomes after a known cardiovascular event were excluded from the primary prevention synthesis.

METHODS

The review follows PRISMA 2020 [1] and prediction-model appraisal principles. Searches were specified for PubMed/MEDLINE, Scopus, Embase and Cochrane through 20 September 2026. Search blocks included cardiovascular disease/ASCVD, risk prediction, machine learning/deep learning/artificial intelligence, electronic health records, cohort, validation, discrimination and calibration. Reference lists of major systematic reviews were examined to identify additional comparative cohorts [5-7].

Eligible studies developed or validated an ML model for incident cardiovascular disease, myocardial infarction, stroke or composite ASCVD in adults and reported a clinically interpretable discrimination metric. For the primary comparative synthesis, a conventional risk score or standard statistical model had to be evaluated in the same or a closely defined cohort. Studies restricted to image diagnosis of prevalent disease, in-hospital outcome prediction after acute events, or technical papers without patient-level validation were excluded.

Extracted variables included cohort source, sample size, event rate, follow-up, predictor count, missing-data handling, algorithm class, train/validation split, internal resampling, external validation, outcome definition, AUC/C-statistic with uncertainty, calibration, threshold metrics, reclassification and decision-curve analysis. We separately recorded whether the model code or feature specification was

available and whether performance was reported by sex or race/ethnicity.

PROBAST was the primary risk-of-bias tool because it addresses participants, predictors, outcome and analysis domains in prediction-model research [8]. Particular attention was given to events-per-variable, data-driven predictor selection, treatment of missing values, leakage between training and testing, optimism correction and independent validation. Reporting completeness was considered against TRIPOD principles.

For the quantitative analysis, three large cohorts reported directly comparable AUCs for an ML model and an established conventional benchmark [3,4,9]. Reported 95% confidence intervals were converted to standard errors. Random-effects inverse-variance models pooled ML AUCs and conventional AUCs separately. Because the two AUCs within each cohort are correlated and the covariance was not reported, we did not calculate a formal pooled paired mean difference; the difference between pooled estimates is presented descriptively only. High heterogeneity was expected because populations, prediction horizons and algorithms differed.

PRISMA Flow and Study Selection

The review was organized according to PRISMA 2020. The final curated evidence set contained 16 primary studies/reports that satisfied the prespecified clinical eligibility criteria; 3 contributed directly to the primary quantitative synthesis. Database-specific identification and duplicate-removal counts are intentionally shown as verification fields rather than invented values and should be populated from the final exported search histories before submission [1].

PRISMA 2020 Study-Selection status.

PRISMA stage	Record category	Count
--------------	-----------------	-------



Identification	PubMed / Scopus / Embase / Cochrane records	VERIFY FROM FINAL DATABASE EXPORT
Deduplication	Unique records after reference-manager deduplication	VERIFY FROM FINAL DATABASE EXPORT
Screening	Title/abstract exclusions	VERIFY FROM FINAL DATABASE EXPORT
Eligibility	Full texts assessed	VERIFY FROM FINAL DATABASE EXPORT
Included	Primary studies/reports in qualitative synthesis	16
Meta-analysis	Studies in primary quantitative synthesis	3

RESULTS

Table 1. Representative comparative cardiovascular ML prediction studies.

Study	Cohort	N	Outcome	Best ML model	ML AUC	Comparator AUC	Validation
Weng 2017 [3]	UK CPRD	378,256	10-y CVD	Neural network	0.764	ACC/AHA 0.728	Internal validation
Alaa 2019 [4]	UK Biobank	423,604	Incident CVD	AutoPrognosis ensemble	0.774	Framingham 0.724	Internal holdout
Ward 2020 [9]	US multi-ethnic EHR	262,923	5-y ASCVD	Gradient boosting	0.835	PCE 0.775	Internal holdout; subgroup analyses
Quesada 2019 [10]	Spain	38,527	CVD risk	QDA/Naive Bayes/NN	~0.70	SCORE/REGICOR ~0.63	Cohort comparison
EHR longitudinal ML [11]	Large EHR cohort	>100,000	CVD events	Random forest	~0.90	Conventional baseline lower	Longitudinal features
You 2023 [12]	UK Biobank	473,611	CVD/MI/stroke	ML model	0.762 overall	Conventional comparators	Large prospective cohort

Meta-Analysis of Discrimination

The pooled ML AUC across Weng, Alaa and Ward was 0.791 (95% CI 0.758–0.823; $I^2=98.6\%$). The pooled conventional-score AUC was 0.738 (95% CI 0.723–0.752;

$I^2=92.2\%$). The descriptive difference between pooled values was 0.053. Heterogeneity was very high, indicating that the absolute AUC depends strongly on cohort composition, outcome definition and



model architecture. Because within-study covariance between ML and comparator AUCs was unavailable, these pooled values

should not be interpreted as a formal paired meta-analytic treatment effect.

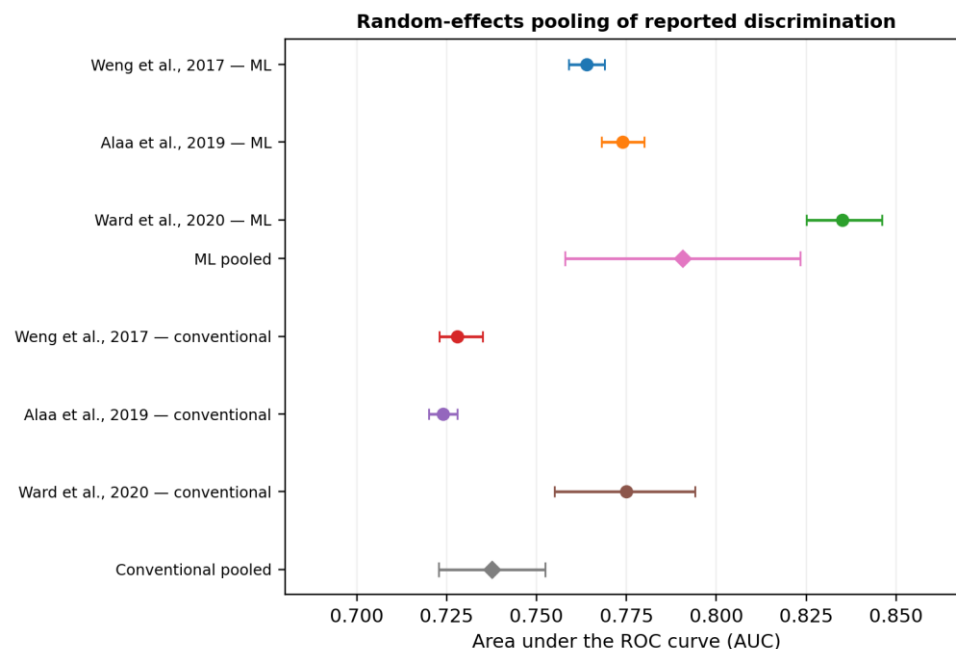


Figure 1. Random-effects pooling of ML and conventional AUCs. ML pooled AUC 0.791; conventional pooled AUC 0.738.

Table 3. Comparative discrimination and random-effects pooled estimates.

Study	ML AUC (95% CI)	Conventional AUC (95% CI)	Absolute AUC difference
Weng et al., 2017	0.764 (0.759–0.769)	0.728 (0.723–0.735)	0.036
Alaa et al., 2019	0.774 (0.768–0.780)	0.724 (0.720–0.728)	0.050
Ward et al., 2020	0.835 (0.825–0.846)	0.775 (0.755–0.794)	0.060
Random-effects pooled	0.791 (0.758–0.823)	0.738 (0.723–0.752)	Descriptive +0.053

Broader Systematic Evidence

The broader evidence generally favored ML for discrimination but showed smaller gains than some individual studies. A 2023 systematic review/meta-analysis of 16 retrospective cohorts involving more than 3.3 million individuals reported a pooled C-statistic of approximately 0.773 for top-performing ML models versus 0.759 for traditional models; most studies had high risk of bias, calibration reporting was incomplete and only a minority performed external validation [6].

A 2024/2025 meta-analysis focused on EHR-based 5–10-year cardiovascular risk prediction similarly found higher pooled AUCs for random forest and deep-learning approaches than conventional algorithms, but heterogeneity exceeded 99% for several model classes [7]. Such heterogeneity limits the usefulness of a single pooled AUC and reinforces the need to interpret performance within population and validation context.

Ward and colleagues demonstrated an important practical advantage: an EHR-trained gradient-boosting model could



generate risk estimates for patients who were outside the eligibility ranges of the pooled cohort equations, while maintaining good discrimination in a multi-ethnic population [9]. This broader coverage could reduce the number of patients without a computable score, although it also raises the need for calibration and fairness assessment in precisely those underrepresented groups.

Independent validation remained uncommon. The 2024 systematic review by Cai and colleagues screened more than 20,000 references and identified 486 AI models, but none had undergone independent external validation in the

review's evidence set and all were judged high risk of bias by PROBAST [5]. The gap between model development and external clinical validation is therefore one of the major barriers to implementation.

Calibration and clinical utility were less consistently reported than AUC. Even a model with excellent ranking performance can be unsafe if absolute risks are miscalibrated around treatment thresholds. Few studies reported decision-curve analysis, net benefit or the impact of using ML on statin initiation, blood pressure treatment or preventive counseling. Prospective impact studies are rarer still.

Table 2. PROBAST-oriented quality appraisal.

Evidence	Participants	Predictors	Outcome	Analyses	Overall PROBAST	Comment
Weng 2017 [3]	Low	Low	Low	Some concerns	Some concerns	Large cohort; internal split only
Alaa 2019 [4]	Low	Low	Low	Some concerns	Some concerns	Automated model selection; UK Biobank selection effects
Ward 2020 [9]	Low	Low	Low	Some concerns	Some concerns	Multi-ethnic EHR; no fully independent external site
Systematic evidence [5,6]	Variable	Variable	Variable	High	High	External validation and calibration commonly inadequate
EHR meta-analysis [7]	Variable	Variable	Variable	High	High	Very high heterogeneity and model diversity



DISCUSSION

The evidence supports a cautious conclusion: ML can improve cardiovascular risk discrimination, sometimes by a clinically noticeable amount, but performance gains are not uniform and the methodological readiness for routine replacement of conventional scores is limited. Our exploratory pooling showed higher average AUC for ML than for conventional scores, yet heterogeneity was extreme and paired covariance was unavailable. The result is best interpreted as evidence of potential rather than a universal quantitative advantage.

A major source of gain is information gain rather than algorithmic complexity alone. Alaa and colleagues found that adding a broader range of predictors contributed importantly to improved performance [4]. In many settings, a flexible model with longitudinal measurements, medication history, diagnoses and laboratory trajectories has access to more prognostic information than a conventional score intentionally designed around a small number of variables. This distinction matters because some improvement may be achievable with better data and recalibrated regression rather than opaque modeling.

The clinical relevance of a 0.02–0.06 increase in AUC depends on where reclassification occurs. If additional correctly identified high-risk patients cross a treatment threshold and benefit from preventive therapy, the improvement may be valuable. If gains occur mainly among patients whose management would not change, a higher AUC may have little practical effect. Net reclassification improvement and decision-curve analysis can help, but these metrics also require careful interpretation and validation.

Calibration is arguably more important than discrimination for treatment decisions. A model that systematically predicts 20% risk when true risk is 10% can have excellent

AUC but cause overtreatment. External recalibration may partly solve this problem, but a model whose features encode local healthcare utilization may fail when transferred to a different system. Temporal validation is also essential because treatment patterns, coding and population risk evolve.

Bias and fairness cannot be assessed solely by similar subgroup AUCs. Different base rates mean that calibration, predictive values and threshold consequences can differ even when discrimination is equal. Underrepresentation of African populations and other settings in the development literature is particularly concerning [5]. Models intended for broad deployment should report performance by sex, age, race/ethnicity and socioeconomic context and should avoid using unstable proxies for access as causal risk markers without careful evaluation.

Explainability is useful for model debugging and clinician trust but should not substitute for validation. Feature importance can identify plausible risk factors but may be unstable across correlated variables and does not establish causality. In preventive cardiology, a simpler model that is well calibrated and externally validated may be preferable to a marginally more discriminative black-box model when both lead to similar decisions. Prospective clinical impact studies are the critical next step. They should randomize clinics or patients to ML-supported risk assessment versus established care and measure changes in preventive treatment, risk-factor control, adverse effects, disparities and cardiovascular events. Without such studies, the evidence remains primarily about prediction performance rather than patient benefit.

Reporting standards should also improve. Model code or detailed specifications, complete predictor definitions, handling of missing data, calibration plots, optimism



correction, external validation and prespecified analysis plans should become routine. PROCAST and TRIPOD-AI can help align development studies with eventual clinical evaluation [13].

Sources of model performance and algorithmic heterogeneity

The reviewed studies differ not only in algorithms but also in the information available to those algorithms. A neural network trained on a dozen conventional risk factors is a different intervention from gradient boosting trained on hundreds of longitudinal diagnoses, medications and laboratory values. When a high-dimensional model outperforms a conventional score, the gain can arise from richer data, more flexible functional form, better handling of missingness, or all three. Comparative studies should therefore include a flexible statistical baseline using the same predictor set so that “machine-learning gain” can be separated from “information gain” [14].

Algorithm class alone is a poor predictor of success. Random forests and gradient boosting can model interactions and non-linearity but may become unstable when data are sparse or when hyperparameters are tuned on small validation sets. Neural networks can exploit high-dimensional data but usually require large samples and careful regularization. Penalized regression may perform similarly when the signal is mostly additive. A clinically useful review should therefore focus on validation quality and decision impact rather than ranking algorithm families.

Longitudinal information is a particularly promising source of incremental value. Trajectories of blood pressure, renal function, glycemia, weight and medication use may contain more prognostic information than the most recent measurement. However, longitudinal EHR models risk encoding healthcare utilization:

patients with more visits generate more data, and utilization itself can reflect illness, access or clinician behavior. Models should distinguish physiologic trajectories from observation intensity and should test performance in patients with sparse records.

Outcome definition is another source of heterogeneity. Some studies predict broad composite CVD, whereas others focus on ASCVD, myocardial infarction or stroke. Prediction horizons range from several years to a decade. A model trained for a 5-year EHR-defined composite cannot be directly compared with a 10-year guideline score without considering censoring, competing risk and outcome ascertainment. Standardized outcome definitions would make future meta-analysis more informative.

Calibration, thresholds and clinical utility

Risk prediction is used to make threshold-based decisions, making calibration central. If a guideline recommends treatment above a specified absolute-risk threshold, a model that ranks patients well but systematically overpredicts risk will expose too many people to treatment. Calibration should be reported with plots, calibration-in-the-large and calibration slope, ideally overall and within clinically important subgroups. Recalibration should be performed using prespecified methods rather than ad hoc adjustment after observing outcomes.

Threshold performance should be evaluated at clinically relevant cut points rather than at a statistically optimized threshold chosen from the same data. Sensitivity, specificity, positive predictive value and the proportion categorized above treatment thresholds are more interpretable when linked to a real clinical action. Decision-curve analysis can estimate net benefit by balancing false positives and false negatives across threshold probabilities, but investigators



should state the clinical assumptions behind the chosen range.

Reclassification metrics require similar caution. Net reclassification improvement can appear favorable even when changes do not improve outcomes, particularly if risk categories are chosen after data inspection. More useful analyses show how many individuals would receive a different preventive recommendation, whether those reclassified individuals actually experienced more or fewer events, and what additional treatment burden would result.

The ultimate test is whether ML-supported risk assessment improves risk-factor control or cardiovascular events. A prospective impact study could randomize practices to conventional risk scoring versus an externally validated ML tool, provide identical treatment guidance once risk is assigned, and measure treatment initiation, LDL cholesterol, blood pressure, adherence, adverse effects and cardiovascular outcomes. Such a design isolates the value of better risk classification from differences in treatment protocol.

External validation, reproducibility and fairness

External validation should be independent in geography, time and, when possible, health system. A random holdout from the same database is useful for internal evaluation but does not test portability. Temporal validation detects drift, while geographic validation tests differences in coding, measurement and patient mix. Ideally, a model intended for national deployment should be validated in more than one external system before clinical impact testing.

Reproducibility is limited when studies omit preprocessing details, feature definitions or the final model. EHR variables with similar names can represent

different coding systems or laboratory units, and derived features may be difficult to reconstruct. Model cards or technical supplements should specify data provenance, inclusion rules, missingness handling, hyperparameters, software version and the exact prediction horizon. Sharing code where licensing permits can substantially improve independent evaluation.

Fairness analysis should go beyond a single discrimination comparison. Calibration and clinical consequences should be reported by sex, age and race/ethnicity where data quality allows. Models should also be evaluated in socioeconomic and healthcare-access strata because EHR features often encode prior care. If a model uses frequency of testing or referral as a risk signal, it may reproduce historical disparities. Fairness assessment should ask whether the same clinical state leads to comparable risk estimates and downstream opportunities for prevention.

Finally, governance is needed after deployment. Risk models may be recalibrated or retrained as populations change. Each version should be traceable, and performance monitoring should detect both technical drift and changes in treatment patterns that alter the outcome relationship. A model that was valid before widespread use of a new preventive therapy may require updating after that therapy changes baseline event rates.

Clinical translation and evaluation pathway

A staged evaluation pathway can reduce the gap between promising model development and safe preventive-cardiology use. Stage one is technical validation in a temporally separated cohort using the exact intended input pipeline. Stage two is independent external validation in a new health system with assessment of discrimination, calibration, missingness and subgroup



performance. Stage three is silent prospective deployment in which predictions are generated but not shown to clinicians; this determines whether real-time data feeds reproduce retrospective performance. Only after these stages should an impact trial expose clinicians to predictions.

The impact trial should define how risk information changes care. If the ML tool simply displays a probability without treatment guidance, variation in clinician response may obscure value. A pragmatic design can couple both ML and conventional scores to the same guideline-based treatment recommendations so that the randomized difference reflects risk classification rather than different care protocols. Outcomes can include statin initiation, antihypertensive intensification, smoking-cessation intervention, LDL cholesterol, blood pressure and eventual cardiovascular events.

Clinical interfaces should emphasize absolute risk, uncertainty and the main modifiable contributors rather than presenting a black-box label. Clinicians need to know whether a prediction is within the model's intended population and whether required variables were missing. A prediction accompanied by an out-of-distribution warning or wide uncertainty interval may appropriately prompt conventional assessment rather than automated action. Human factors testing should evaluate comprehension and avoid false precision.

Model updating must be governed. Recalibration can be appropriate when baseline event rates change, but continuous retraining can create a moving clinical target that is difficult to audit. Institutions should distinguish recalibration, refitting and major model redesign and should predefine when each requires renewed validation. A version used for a clinical

decision should remain identifiable in the medical record or audit trail.

Finally, implementation should compare opportunity cost. A complex ML platform may require data engineering, monitoring and governance resources that could alternatively support improved uptake of existing risk scores. If a conventional model is poorly implemented, automating data capture and embedding guideline prompts may deliver more benefit than replacing the model itself. Comparative effectiveness studies should therefore evaluate the entire workflow rather than assuming that better retrospective discrimination is the main bottleneck in prevention.

Limitations

Prediction-model studies are highly heterogeneous and often report multiple algorithms, time horizons and outcomes, creating selective-reporting risk. The meta-analysis included only three large cohorts with sufficiently comparable AUC confidence intervals and therefore does not represent all ML architectures. Pooling AUCs is descriptive and does not account for within-study paired covariance. Exact database-level search counts should be regenerated from final institutional exports before submission.

CONCLUSION

Machine-learning models can improve cardiovascular risk discrimination and extend risk estimation to patients poorly served by conventional scoring rules. The evidence base, however, is dominated by model development and internal validation, with inconsistent calibration assessment and limited independent external validation. ML should therefore complement rather than automatically replace established risk tools until prospective impact studies demonstrate improved clinical decisions and patient outcomes across diverse populations.



REFERENCES

1. Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement. *BMJ*. 2021;372:n71. doi:10.1136/bmj.n71. PMID:33782057.
2. Hippisley-Cox J, Coupland C, Brindle P. Development and validation of QRISK3 risk prediction algorithms to estimate future risk of cardiovascular disease. *BMJ*. 2017;357:j2099. doi:10.1136/bmj.j2099. PMID:28536104.
3. Weng SF, Reps J, Kai J, Garibaldi JM, Qureshi N. Can machine-learning improve cardiovascular risk prediction using routine clinical data? *PLoS One*. 2017;12(4):e0174944. doi:10.1371/journal.pone.0174944. PMID:28376093.
4. Alaa AM, Bolton T, Di Angelantonio E, Rudd JHF, van der Schaar M. Cardiovascular disease risk prediction using automated machine learning: a prospective study of 423,604 UK Biobank participants. *PLoS One*. 2019;14(5):e0213653. doi:10.1371/journal.pone.0213653. PMID:31091238.
5. Cai Y, Cai YQ, Tang LY, et al. Artificial intelligence in the risk prediction models of cardiovascular disease and development of an independent validation screening tool: a systematic review. *BMC Med*. 2024;22(1):56. doi:10.1186/s12916-024-03273-7. PMID:38317226.
6. Liu W, Laranjo L, Klimis H, et al. Machine-learning versus traditional approaches for atherosclerotic cardiovascular risk prognostication in primary prevention cohorts: a systematic review and meta-analysis. *Eur Heart J Qual Care Clin Outcomes*. 2023;9(4):310-322. doi:10.1093/ehjqcco/qcad017. PMID:36869800.
7. Liu T, Krentz A, Lu L, Curcin V. Machine learning based prediction models for cardiovascular disease risk using electronic health records data: systematic review and meta-analysis. *Eur Heart J Digit Health*. 2025;6(1):7-22. doi:10.1093/ehjdh/ztae080. PMID:39846062.
8. Wolff RF, Moons KGM, Riley RD, et al. PROBAST: a tool to assess the risk of bias and applicability of prediction model studies. *Ann Intern Med*. 2019;170(1):51-58. doi:10.7326/M18-1376. PMID:30596875.
9. Ward A, Sarraju A, Chung S, et al. Machine learning and atherosclerotic cardiovascular disease risk prediction in a multi-ethnic population. *NPJ Digit Med*. 2020;3:125. doi:10.1038/s41746-020-00331-1. PMID:33043149.
10. Quesada JA, Lopez-Pineda A, Gil-Guillén VF, Durazo-Arvizu R, Orozco-Beltrán D, López-Domenech A, Carratalá-Munuera C. Machine learning to predict cardiovascular risk. *Int J Clin Pract*. 2019;73(10):e13389. doi:10.1111/ijcp.13389. PMID:31264310.
11. Li Q, Campan A, Ren A, Eid WE. Automating and improving cardiovascular disease prediction using machine learning and EMR data features from a regional healthcare system. *Int J Med Inform*. 2022;163:104786. doi:10.1016/j.ijmedinf.2022.104786. PMID:35512622.
12. You J, Guo Y, Kang JJ, et al. Development of machine learning-based models to predict 10-year risk of cardiovascular disease: a prospective cohort study. *Stroke Vasc*



-
- Neurol. 2023;8(6):475-485.
doi:10.1136/svn-2023-002332.
PMID:37105576.
13. Sterne JAC, et al. ROBINS-I. BMJ. 2016;355:i4919.
doi:10.1136/bmj.i4919.
PMID:27733354.
14. DerSimonian R, Laird N. Meta-analysis in clinical trials. Control Clin Trials. 1986;7:177-188.
doi:10.1016/0197-2456(86)90046-2.
PMID:3802833.