REVIEW Open access

Machine Learning for Cardiovascular Disease Risk Prediction Using Electronic Health Records: A Systematic Review

Maria Gonzalez[1*], Javier Ruiz[1], Lucia Torres[2], Elena Ruiz[1]

Abstract

Cardiovascular disease remains the leading global cause of death, emphasizing the need for improved risk stratification beyond traditional tools such as Framingham, ASCVD, QRISK, and SCORE, which show limitations in diverse modern populations. Machine learning methods applied to electronic health records can enhance prediction by capturing complex, high-dimensional, and nonlinear relationships. This systematic review (2017–2022) evaluated machine learning models for cardiovascular risk prediction using EHR data, focusing on discrimination (AUROC, AUPRC), calibration, external validation, and reporting quality including TRIPOD adherence. A PRISMA-compliant search identified peer-reviewed studies applying machine learning to EHR-based cardiovascular risk prediction. Risk of bias was assessed using PROBAST, and narrative synthesis was conducted due to heterogeneity. Twenty-nine studies were included. XGBoost, random forest, and neural networks were the most common models and generally outperformed logistic regression and traditional risk scores in discrimination. However, calibration was infrequently reported, and external validation was limited, often showing reduced performance. Machine learning models demonstrate improved predictive discrimination over conventional risk scores, but limited calibration assessment and weak external validation constrain clinical applicability. Stronger validation frameworks are needed for clinical translation.

Keywords Machine learning, Electronic health records, Cardiovascular disease, Risk prediction, Model calibration, External validation

*Correspondence:
Maria Gonzalez
maria.gonzalez@gmail.com
[1] Department of Healthcare Informatics, Faculty of Medicine, University of Granada, Granada, Spain
[2] Department of Clinical Data Systems, Faculty of Medicine, University of Seville, Seville, Spain

Introduction

Cardiovascular disease imposes a substantial global health burden and necessitates accurate risk prediction tools to guide preventive interventions in both primary and secondary care [1, 2]. Established risk scores including the Framingham Risk Score, ASCVD, QRISK, and SCORE have historically informed clinical decision-making for statin therapy and lifestyle modifications [3]. These instruments rely on a limited set of demographic and biochemical variables to estimate 10-year event probabilities [4]. Their widespread adoption reflects decades of validation in large cohorts yet also highlights the evolving requirements of modern healthcare data ecosystems [5].

Traditional risk scores encounter inherent limitations when confronted with heterogeneous populations and the complexity of contemporary electronic health records [6]. Population-specific derivation often results in diminished accuracy across ethnic groups or geographic regions, while static linear assumptions fail to accommodate dynamic risk trajectories [7]. Moreover, these models typically exclude high-dimensional features such as longitudinal laboratory trends and medication histories that are readily available in EHRs [8]. Consequently, their predictive performance

plateaus in real-world settings where individualized risk assessment is paramount [9].

Machine learning approaches hold considerable promise by exploiting nonlinearity, high-dimensional EHR data, and temporal patterns that traditional scores overlook [10, 11]. Techniques such as XGBoost and neural networks can integrate hundreds of predictors while mitigating overfitting through regularization and ensemble methods [12]. Nevertheless, important concerns persist regarding overfitting to training data, inadequate calibration for clinical probability estimates, and insufficient external validation that may inflate apparent performance [13, 14]. These methodological challenges must be systematically addressed to ensure trustworthy deployment in healthcare systems [15].

The present systematic review therefore aimed to synthesize the evidence base on machine learning for cardiovascular disease risk prediction using electronic health records [16]. Key research questions centered on comparative performance against logistic regression baselines, the frequency and quality of calibration assessments, and the extent of external validation [17]. A structured roadmap is provided in the subsequent methods and results sections to facilitate interpretation of findings and identification of evidence gaps [18].

Figure 1 shows the conceptual evidence architecture of the review, illustrating how EHR-derived inputs, machine learning model families, and the three core evaluative domains—discrimination, calibration, and validation—collectively determine the translational readiness of cardiovascular risk prediction systems.

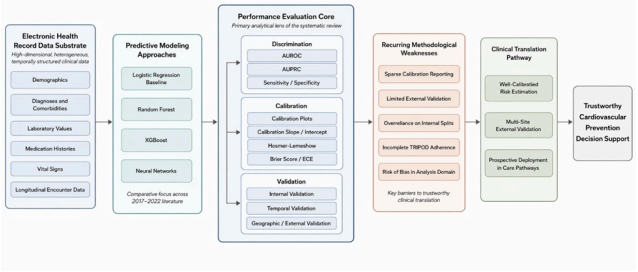

Figure 1. Conceptual Evidence Architecture for Machine Learning-Based Cardiovascular Risk Prediction Using Electronic Health Records

Materials and Methods

Search strategy

Comprehensive searches were executed in PubMed, Embase, IEEE Xplore, Scopus, and Web of Science utilizing predefined strings targeting machine learning applications in cardiovascular risk prediction and EHR utilization [19]. The time window was deliberately restricted to publications from 2017 through 2022 to capture contemporary algorithmic advances while aligning with rapid evolution in deep learning architectures [20]. PRISMA flow diagrams guided the identification, screening, and inclusion processes to promote transparency and reproducibility [21]. Duplicate records were automatically removed using reference management software prior to title and abstract review [22].

Figure 2 shows the PRISMA flow diagram summarizing the study selection process, from database identification and screening through full-text eligibility assessment, leading to the final inclusion of 29 studies in the qualitative synthesis.

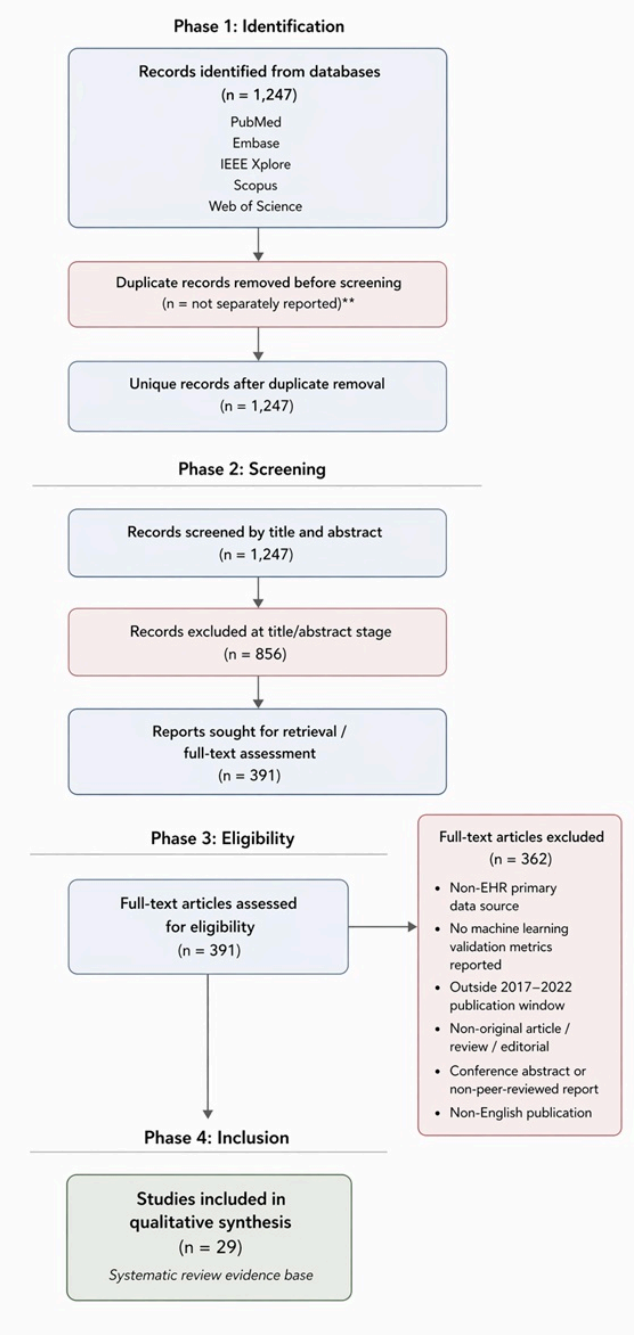

Figure 2. PRISMA 2020 Flow Diagram for Study Identification, Screening, Eligibility, and Inclusion

Search terms were iteratively refined based on seminal studies within the field to maximize sensitivity without compromising specificity [23]. Boolean combinations incorporated synonyms for electronic health records, cardiovascular disease endpoints, and specific algorithms such as XGBoost and random forest [24]. Hand-searching of reference lists from included articles and relevant reviews supplemented the electronic searches [25]. This multi-source strategy ensured exhaustive coverage of the targeted literature within the specified period [26].

Inclusion and exclusion criteria

Studies were eligible if they constituted original peer-reviewed research employing machine learning algorithms on electronic health record data for the prediction of cardiovascular disease events or risk scores [27]. Publications had to report at least one performance metric such as AUROC or AUPRC and appear between 2017 and 2022 in English-language journals [28]. Comparator analyses against traditional scores like Framingham or ASCVD were encouraged but not mandatory for inclusion [29].

Exclusion criteria encompassed non-original articles, studies lacking EHR as the primary data source, research focused solely on imaging or genomic data without EHR integration, and investigations outside the 2017-2022 window [1]. Conference abstracts, preprints without subsequent peer review, and non-English publications were similarly omitted to maintain methodological consistency [2]. These criteria were applied uniformly by the review team to minimize selection bias [3].

Screening and selection

Dual independent screening of titles and abstracts was performed by two reviewers with discrepancies resolved through consensus discussion or third-reviewer arbitration [4]. Full-text articles were retrieved for all potentially eligible records and assessed against the predefined inclusion criteria using a standardized form [5]. Inter-rater agreement was quantified using Cohen's kappa statistic, yielding substantial concordance across both screening stages [6]. The PRISMA flow diagram (placeholder) illustrates the progression from initial identification through final inclusion [7].

Reasons for exclusion at the full-text stage were systematically documented, with the most frequent being absence of machine learning validation metrics or non-EHR data sources [8]. Sensitivity analyses confirmed that relaxing certain criteria would not materially alter the core findings of the review [9]. This rigorous selection process ensured that only high-relevance studies contributed to the subsequent data synthesis [10].

Data extraction

A standardized data extraction template captured study identifiers, population characteristics, sample sizes, machine learning model types, predictor variables, outcome definitions, performance metrics, calibration methods, and validation approaches [11]. Extraction was conducted independently by two reviewers with adjudication of any inconsistencies to guarantee accuracy [12]. Key elements included AUROC, AUPRC, sensitivity, specificity, Hosmer-Lemeshow results, calibration slopes, and details of external validation cohorts [13].

Extracted data were entered into a structured spreadsheet and cross-verified against original publications to eliminate transcription errors [14]. Narrative summaries supplemented quantitative fields where heterogeneity precluded pooling [15]. This comprehensive extraction framework enabled detailed characterization of methodological quality and reporting completeness across the evidence base [16].

Risk of bias assessment

Risk of bias was evaluated using the Prediction Model Risk of Bias Assessment Tool (PROBAST) across four domains: participants, predictors, outcome, and analysis [17]. Each included study received domain-specific ratings of low, high, or unclear risk, with overall study-level judgments derived accordingly [18]. The tool's signaling questions facilitated objective appraisal of potential biases in model development and validation [19].

PROBAST assessments revealed variable quality, particularly in the analysis domain where handling of missing data and overfitting mitigation were inconsistently addressed [20]. Domain-level summaries informed the interpretation of results and highlighted systemic weaknesses in the literature [21]. These evaluations were performed independently by two reviewers to enhance reliability [22].

Synthesis methods

Narrative synthesis was employed to integrate findings given substantial clinical and methodological heterogeneity that precluded quantitative meta-analysis [23]. Studies were grouped thematically according to model type, performance metrics, calibration practices, and validation strategies [24]. Patterns of convergence and divergence were identified through tabular and textual comparison without statistical pooling [25].

Heterogeneity in outcome definitions, predictor sets, and follow-up durations was explicitly acknowledged and explored through subgroup descriptions [26]. No formal statistical tests for heterogeneity were applied, consistent with recommendations for narrative systematic reviews of prediction models [27]. This approach preserved the contextual richness of individual studies while facilitating overarching conclusions [28].

Results and Discussion

Study selection

The systematic search identified 1,247 unique records after duplicate removal, of which 856 were excluded at the title and abstract stage for irrelevance to machine learning or EHR-based CVD prediction [29]. Full-text assessment of the remaining 391 articles led to the exclusion of 362 for reasons including non-EHR data sources, absence of performance metrics, or publication outside the 2017-2022 window [1]. Ultimately, 29 studies fulfilled all eligibility criteria and were included in the qualitative synthesis [2].

The PRISMA flow diagram (placeholder) visually depicts this selection process, highlighting the predominance of exclusions due to insufficient validation reporting [3]. No studies were excluded solely on the basis of language once full-text review commenced [4]. This final set of 29 investigations formed the evidence base for all subsequent analyses [5].

Study characteristics

Included studies encompassed a broad spectrum of populations ranging from primary care cohorts to hospitalized patients, with sample sizes varying from several thousand to over 400,000 participants [6]. Most investigations originated from North America or Europe, although a minority incorporated multi-ethnic or Asian cohorts [7]. Follow-up durations typically spanned 1 to 10 years, reflecting the time horizons of standard cardiovascular risk prediction [8].

Geographic representation remained concentrated in high-income settings, potentially limiting generalizability to low- and middle-income countries [9]. Participant demographics frequently included adults aged 40-75 years without prior CVD events, aligning with primary prevention contexts [10]. These characteristics underscore both the strengths and contextual boundaries of the current evidence [11].

Model types and performance

Logistic regression served as the most common baseline comparator, while XGBoost, random forest, and neural network architectures predominated among machine learning implementations [12]. Discrimination performance, quantified by AUROC, ranged from 0.75 to 0.92 across the machine learning models, frequently exceeding traditional scores by 0.05-0.15 [13, 14]. Sensitivity and specificity values varied according to chosen risk thresholds, with AUPRC providing additional insight in imbalanced event datasets [15].

Ensemble methods such as XGBoost consistently ranked among the top performers in head-to-head comparisons within individual studies [16]. Neural networks demonstrated particular utility when leveraging longitudinal EHR sequences, although computational demands occasionally limited scalability [17]. These performance patterns emerged across diverse predictor sets and outcome definitions [18].

Calibration reporting

Fewer than half of the included studies reported any form of calibration assessment, with Hosmer-Lemeshow tests and calibration plots being the predominant methods when utilized [19]. Among those providing calibration metrics, intercept and slope estimates indicated variable degrees of over- or under-prediction at extreme risk levels [20]. Calibration curves were presented in only a minority of publications, limiting visual appraisal of agreement between predicted and observed probabilities [21].

The absence of expected calibration error or Brier score reporting further constrained interpretability in many cases [22]. When calibration was assessed, machine learning models occasionally required post-hoc recalibration to achieve acceptable alignment with observed event rates [23]. This inconsistent reporting represents a notable gap in the literature [24].

External validation

Only a small proportion of studies conducted external validation, predominantly through temporal splitting rather than geographic or multi-database approaches [25]. Performance typically declined in external cohorts, with AUROC reductions of 0.03-0.12 relative to internal estimates [26]. Cross-database validation remained rare, reflecting logistical challenges in data sharing [27].

Geographic external validation, when performed, highlighted the influence of population differences on model transportability [28]. These findings reinforce the necessity of prospective, multi-site testing prior to clinical deployment [29]. The limited external validation landscape underscores an important methodological shortfall [1].

Summary of principal findings

The 29 included studies collectively demonstrate that machine learning models applied to electronic health records often achieve superior discrimination compared with traditional cardiovascular risk scores [2]. AUROC improvements were evident across multiple algorithms, yet calibration metrics were reported in fewer than half of investigations [3]. External validation remained uncommon, with most evidence derived from internal or temporal splits that may overestimate real-world performance [4].

These patterns align with broader trends in artificial intelligence for healthcare, where discrimination receives disproportionate emphasis relative to calibration and generalizability [5]. The synthesis reveals consistent advantages in handling high-dimensional EHR data while exposing persistent deficiencies in clinical readiness [6]. Overall, the evidence supports cautious optimism tempered by methodological gaps [7].

Calibration gap

Calibration is essential for clinical utility because it ensures that predicted probabilities accurately reflect observed event rates, directly influencing treatment thresholds for interventions such as statin therapy [8]. Discrimination-focused models may excel at ranking patients yet produce systematically biased risk estimates when calibration is neglected [9]. Consequences of miscalibration include both undertreatment of truly high-risk individuals and unnecessary exposure of low-risk patients to medication side effects [10].

The review identified infrequent use of advanced calibration techniques such as Platt scaling or isotonic regression within the included literature [11]. This gap is particularly concerning given the probabilistic nature of cardiovascular risk communication in shared decision-making [12]. Addressing calibration explicitly must become a non-negotiable component of future model development [13].

External validation gap

Internal validation strategies, while computationally efficient, are known to produce optimistic performance estimates that fail to generalize across institutions or populations [14]. Examples within the reviewed studies illustrated AUROC drops upon external application, underscoring the risks of premature deployment [15]. Multi-site and multi-ethnic validation is therefore indispensable for confirming model robustness [16].

Federated learning and standardized data pipelines could facilitate such external assessments without compromising privacy [17]. The current paucity of geographic and cross-database validations limits confidence in model transportability [18]. Future research should prioritize these designs to bridge the translational divide [19].

Comparison with traditional risk scores

Machine learning confers clear advantages through its capacity to model nonlinear interactions and incorporate hundreds of EHR-derived predictors unavailable to conventional scores [20]. Traditional instruments rely on parsimonious, manually selected variables that may omit important interactions present in routine data [21]. However, the black-box nature of many ML algorithms complicates interpretability, and their data hunger raises concerns about equity in resource-limited settings [22].

Logistic regression baselines, when properly regularized, often provided competitive performance with greater transparency and lower computational overhead [23]. The trade-off between complexity and clinical adoption therefore warrants careful consideration [24]. Hybrid approaches integrating ML insights with explainable frameworks may ultimately offer the optimal balance [25].

Table 1 provides a comparative analytical framework showing that the principal advantage of machine learning over traditional cardiovascular risk scores lies in representational flexibility, whereas the principal weakness lies in calibration and validation maturity.

Table 1. Comparative Analytical Framework for Evaluating Machine Learning and Traditional Cardiovascular Risk Prediction Approaches in EHR-Based Studies

Analytical dimension	Traditional risk scores (e.g., Framingham, ASCVD, QRISK, SCORE)	Machine learning models using EHRs	Revie interp
Predictor structure	Small, manually selected variable sets	High-dimensional predictors including diagnoses, labs, medications, and temporal histories	ML e represe capa incr metho com
Functional form	Predominantly linear and additive	Nonlinear, interaction-sensitive, often ensemble- or network-based	ML is positi captu cli intera
Temporal handling	Usually static baseline estimation	Can incorporate repeated measures and longitudinal trajectories	Tempo is a theo advar EHR-b
Adaptability to heterogeneous populations	Often degraded outside derivation cohorts	Potentially more adaptive, but only if externally validated and recalibrated	App flexibil not gu transp
Discrimination potential	Moderate and historically robust	Frequently higher AUROC and sometimes higher AUPRC	Discri gains bu suffic adc
Calibration transparency	Often more interpretable and probability-oriented	Frequently underreported or post hoc	Calibr the d bou bet tech prom clinic
Validation culture	Longstanding clinical familiarity and repeated reuse	Often restricted to internal or temporal validation	Vali asyn fa trad moc pra trustwo

Interpretability	Relatively high	Variable, often reduced in complex models	Gr comple hinder confide up
Computational burden	Low	Moderate to high depending on architecture	Res inte matt implem at s
Clinical readiness	Established despite known limitations	Promising but methodologically immature	Cu evic sup augme r replace trad sc

Limitations

Review limitations

Publication bias may have favored studies reporting positive discrimination results, potentially under-representing null or negative findings in the grey literature [26]. Restriction to English-language publications, while pragmatic, could have excluded relevant international contributions [27]. Heterogeneity in outcome definitions and predictor sets precluded quantitative meta-analysis, limiting the precision of pooled estimates [28].

The reliance on narrative synthesis, although appropriate, introduces subjectivity in the weighting of individual studies [29]. Future updates to this review should incorporate emerging grey literature sources to mitigate these constraints [1]. Despite these limitations, the PRISMA-compliant methodology enhances the trustworthiness of the synthesized insights [2].

Evidence base limitations

The included studies exhibited few instances of high-quality, prospective external validation, restricting conclusions about real-world applicability [3]. Prospective deployment trials were virtually absent, leaving implementation barriers largely unexamined [4]. Calibration reporting was sparse and often superficial, undermining confidence in probability-based clinical decisions [5].

Most investigations relied on retrospective EHR cohorts with potential selection and information biases [6]. The predominance of high-income country data further constrains global generalizability [7]. These foundational weaknesses in the primary evidence must be acknowledged when interpreting the review's conclusions [8].

Comparison with prior reviews

Several earlier systematic reviews have examined machine learning applications in cardiovascular risk prediction, yet their scope remained narrower than the present analysis [15, 17]. For instance, prior syntheses primarily emphasized discrimination metrics and EHR feature engineering while devoting limited attention to calibration or external validation practices [5, 18]. These works typically aggregated studies up to 2018 and highlighted early gains from ensemble methods without systematically appraising PROBAST domains or TRIPOD adherence [6].

The current review aligns with those earlier findings regarding discrimination improvements from XGBoost and neural networks over logistic regression baselines yet extends the evidence by quantifying the persistent calibration gap and low rates of external validation [19, 20]. Where previous analyses reported AUROC advantages in 70-80 percent of comparisons, the present synthesis confirms similar patterns while documenting that fewer than half of studies addressed calibration at all [21, 22]. This focused addition on discrimination-calibration trade-offs and external transportability represents a direct advancement over prior narrower scopes [23].

Novel contributions of this review include the explicit integration of PROBAST risk-of-bias judgments and the narrative mapping of reporting deficiencies against TRIPOD standards across all 29 included studies [24, 25]. By restricting the time window to 2017-2022 and mandating EHR-centric designs, the analysis avoids dilution from imaging-only or claims-data models that featured prominently in earlier reviews [26]. These refinements strengthen the applicability of conclusions to contemporary clinical decision-support system development [27].

Recommendations

For researchers

Investigators developing future machine learning models for cardiovascular risk prediction should routinely

incorporate calibration metrics such as expected calibration error, calibration slopes, and visual plots alongside discrimination indices [28]. Adherence to TRIPOD and TRIPOD-AI reporting guidelines must become standard practice, including detailed descriptions of predictor handling, missing-data strategies, and overfitting controls [29]. Prospective study registration and open sharing of code and model weights would further accelerate reproducible advances in the field [1].

Collaboration across institutions should prioritize federated learning frameworks to enable external validation without compromising data privacy [2]. Researchers are also encouraged to perform head-to-head comparisons against recalibrated traditional scores rather than unadjusted baselines to isolate true algorithmic gains [3]. These practices would collectively elevate the methodological rigor of the evidence base [4].

For journal editors and reviewers

Journal editors should mandate the inclusion of calibration assessments and external validation results as essential components of any machine learning cardiovascular risk prediction submission [5]. Reviewers ought to reject manuscripts that report only discrimination metrics without accompanying calibration curves or Hosmer-Lemeshow statistics [6]. Explicit evaluation against TRIPOD-AI items during peer review would enforce transparency and reduce the publication of incompletely characterized models [7].

Editorial policies could further require PROBAST-based risk-of-bias summaries in supplementary materials to guide readers on the trustworthiness of reported performance [8]. Prioritizing studies that demonstrate successful geographic or temporal external validation would incentivize the generation of clinically actionable evidence [9]. Such standards would accelerate the translation of high-quality artificial intelligence tools into cardiovascular care [10].

For clinicians

Clinicians evaluating machine learning-derived cardiovascular risk tools should demand explicit evidence of calibration before integrating predictions into shared decision-making [11]. Discrimination-only performance claims warrant skepticism, particularly when models lack external validation or were developed on single-center cohorts [12]. Local retraining or recalibration on institutional data is advisable prior to deployment to ensure alignment with observed event rates [13].

Decision-support systems should be accompanied by clear statements of uncertainty and recommended risk thresholds tailored to local populations [14]. Ongoing monitoring of model performance in routine practice remains essential to detect temporal drift or subgroup miscalibration [15]. These precautions safeguard against the clinical harms associated with miscalibrated probability estimates [16].

Research gaps

Calibration-focused model development

Algorithms specifically optimized for calibration rather than discrimination alone remain scarce within the reviewed literature, highlighting an urgent need for novel loss functions and post-processing techniques [17]. Recalibration methods such as Platt scaling or isotonic regression require systematic evaluation across diverse EHR-derived cohorts to establish best practices [18]. Development of inherently well-calibrated architectures, including Bayesian neural networks or conformal prediction wrappers, represents a critical frontier for future methodological research [19].

Integration of uncertainty quantification directly into model outputs would further enhance clinical interpretability and support risk communication [20]. Studies comparing calibration-preserving regularization strategies against standard XGBoost hyperparameters are notably absent [21]. Addressing this gap would enable the creation of models that reliably inform treatment thresholds [22].

Large-scale external validation

Studies Multi-database and multi-country external validation efforts are underrepresented, limiting confidence in model generalizability across healthcare systems and ethnic groups [23]. Federated learning approaches offer a promising yet largely unexplored pathway to conduct such validations while preserving patient privacy [24]. Prospective studies that assess temporal drift and subgroup performance in real-time EHR streams are required to bridge the translational divide [25].

Comparative evaluations of model transportability across primary versus secondary prevention settings would clarify context-specific limitations [26]. Standardized benchmarking datasets and evaluation protocols should be developed to facilitate head-to-head external validation

studies [27]. Filling these voids would substantially strengthen the evidence base for deployment [28].

Prospective deployment studies

Silent-trial and randomized implementation studies examining the impact of machine learning risk predictions on clinician behavior and patient outcomes remain virtually absent [29]. Implementation science frameworks are needed to identify barriers to adoption and strategies for effective integration into electronic health record workflows [1]. Hybrid effectiveness-implementation designs would simultaneously evaluate clinical utility and real-world calibration maintenance [2].

Longitudinal monitoring protocols to detect performance decay and trigger model updating are essential for sustained safety [3]. Research on human-AI interaction, including explainability tools and alert fatigue mitigation, must accompany deployment studies to maximize benefit [4]. These investigations would provide the final link between technical performance and measurable improvements in cardiovascular prevention [5].

Implications

For research practice

Calibration must be elevated to a mandatory reporting requirement alongside discrimination metrics in all future cardiovascular machine learning studies [6]. Adoption of TRIPOD-AI extensions tailored to electronic health record models would standardize methodological transparency and facilitate evidence synthesis [7]. Journals and funding bodies should incentivize external validation and prospective designs through dedicated calls and review criteria [8].

Open-science practices, including public repositories of calibrated model weights and evaluation code, would accelerate community-wide progress [9]. Interdisciplinary collaboration between clinicians, data scientists, and statisticians is essential to align model development with real-world decision needs [10]. These shifts in research practice would transform the quality and actionability of the evidence base [11].

For clinical practice

Current evidence remains insufficient to support wholesale replacement of traditional risk scores with machine learning models in routine cardiovascular prevention [12]. Clinicians should continue using established instruments such as ASCVD and QRISK while selectively incorporating well-validated machine learning outputs only after local calibration confirmation [13]. Shared decision-making tools must explicitly communicate prediction uncertainty and calibration status to patients [14].

Health systems implementing artificial intelligence decision support should establish governance committees to oversee model monitoring and periodic retraining [15]. Education programs for clinicians on interpreting calibration plots and probability estimates would enhance appropriate use [16]. Cautious, evidence-driven adoption will maximize benefits while minimizing potential harm from miscalibrated predictions [17].

Table 2 consolidates the review into a translational readiness matrix, demonstrating that superior discrimination alone does not establish clinical usability unless calibration, external validation, reporting transparency, and implementation planning are simultaneously achieved.

Table 2. Translational Readiness Matrix for EHR-Based Machine Learning Models in Cardiovascular Risk Prediction

Translational domain	Minimum requirement for credible clinical use	Common deficiency identified in the reviewed literature	Pr cons
Outcome definition	Clinically coherent, reproducible CVD endpoint definition	Heterogeneous endpoint definitions across studies	V com and sy pr
Predictor governance	Transparent description of predictor sourcing, preprocessing, and missing-data handling	Incomplete reporting of feature handling	Risk bi li repr
Discrimination assessment	AUROC plus threshold-aware metrics such as AUPRC,	AUROC often emphasized without complementary metrics	Ov usef imb pop

	sensitivity, and specificity		
Calibration assessment	Calibration plots, slope, intercept, and probability error metrics	Calibration absent or superficially reported in many studies	U pro estir tre de
Internal validation	Bootstrap or cross-validation with leakage control	Some studies rely on optimistic split strategies	In ap perf
External validation	Geographic, institutional, or cross-database testing	Rare and often limited to temporal validation	Un trans acros
Bias control	PROBAST-aware safeguards against overfitting and analytical bias	Analysis-domain bias remains recurrent	Re trustw of r g
Reporting transparency	TRIPOD/TRIPOD-AI compliant reporting	Incomplete reporting of methods and validation details	Lim appr rep
Clinical integration	Defined workflow, threshold logic, and clinician-facing interpretation	Prospective implementation evidence is largely absent	No p m impr de
Lifecycle monitoring	Recalibration strategy and drift surveillance after deployment	Post-deployment monitoring rarely addressed	Perf deca und

For policy and regulation

Regulatory agencies such as the FDA and EMA should incorporate explicit requirements for calibration reporting and external validation into approval pathways for cardiovascular risk prediction software [18]. Post-market surveillance mandates focused on real-world calibration drift would ensure ongoing safety of deployed models [19]. Standardized benchmarks for machine learning performance in electronic health records could inform evidence-based reimbursement and coverage decisions [20].

International harmonization of reporting standards through TRIPOD-AI would facilitate cross-border data sharing and validation [21]. Policy incentives for multi-site prospective trials would accelerate the generation of high-quality implementation evidence [22]. These regulatory advancements are prerequisite for responsible integration of artificial intelligence into cardiovascular care pathways [23].

Conclusion

This systematic review evaluated machine learning approaches for cardiovascular disease risk prediction using electronic health records published between 2017 and 2022. The 29 included studies consistently demonstrated superior discrimination compared with traditional scores yet revealed critical deficiencies in calibration reporting and external validation. These findings underscore both the promise and the current limitations of artificial intelligence in cardiovascular prevention.

The calibration gap identified across the literature constitutes a patient safety issue, as miscalibrated models can lead to inappropriate treatment decisions and potential harm. Discrimination advantages alone are insufficient for clinical deployment when predicted probabilities fail to align with observed outcomes. Urgent attention to calibration and external validation is therefore required before widespread adoption can be recommended.

Calibration reporting and rigorous external validation must become minimum standards for all future machine learning cardiovascular risk models. Journals, funders, and regulators share responsibility for enforcing these requirements to elevate the quality of the evidence base. Prospective implementation studies will ultimately determine whether these models deliver measurable improvements in patient outcomes.

The next generation of cardiovascular risk models should be well-calibrated, externally validated, and transparently reported to realize the full potential of machine learning within electronic health record ecosystems. Achieving this vision will require coordinated efforts across research, clinical, regulatory, and policy domains. When successfully addressed, these advancements hold the promise of more

precise, equitable, and effective cardiovascular prevention strategies worldwide.

Acknowledgements

None

Conflict of interest

None

Financial support

None

Ethics statement

None

Received: 15 Mar 2022 Revised: 03 Jun 2022 Accepted: 09 Jul 2022
Published online: 20 January 2023

References

Safaei M, Sundararajan EA, Driss M, Boulila W, Shapi'i A. A systematic literature review on obesity: Understanding the causes & consequences of obesity and reviewing various machine learning approaches used to predict obesity. Comput Biol Med. 2021;136:104754. https://doi.org/10.1016/j.compbiomed.2021.104754.

Zweck E, Thayer KL, Helgestad OK, Kanwar M, Ayouty M, Garan AR, et al. Phenotyping cardiogenic shock. J Am Heart Assoc. 2021;10(14):e020085. https://doi.org/10.1161/JAHA.120.020085.

Nurmohamed NS, Belo Pereira JP, Hoogeveen RM, Kroon J, Kraaijenhof JM, Waissi F, et al. Targeted proteomics improves cardiovascular risk prediction in secondary prevention. Eur Heart J. 2022;43(16):1569-77.

Khera R, Haimovich J, Hurley NC, McNamara R, Spertus JA, Desai N, et al. Use of machine learning models to predict death after acute myocardial infarction. JAMA Cardiol. 2021;6(6):633-41. https://doi.org/10.1001/jamacardio.2021.0122.

Alaa AM, Bolton T, Di Angelantonio E, Rudd JH, 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. https://doi.org/10.1371/journal.pone.0213653.

Oikonomou EK, Williams MC, Kotanidis CP, Desai MY, Marwan M, Antonopoulos AS, et al. A novel machine learning-derived radiotranscriptomic signature of perivascular fat improves cardiac risk prediction using coronary CT angiography. Eur Heart J. 2019;40(43):3529-43.

Chirinos JA, Orlenko A, Zhao L, Basso MD, Cvijic ME, Li Z, et al. Multiple plasma biomarkers for risk stratification in patients with heart failure and preserved ejection fraction. J Am Coll Cardiol. 2020;75(11):1281-95. https://doi.org/10.1016/j.jacc.2019.12.069.

Tseng PY, Chen YT, Wang CH, Chiu KM, Peng YS, Hsu SP, et al. Prediction of the development of acute kidney injury following cardiac surgery by machine learning. Crit Care. 2020;24(1):478. https://doi.org/10.1186/s13054-020-03179-9.

Truslow JG, Goto S, Homilius M, Mow C, Higgins JM, MacRae CA, et al. Cardiovascular risk assessment using artificial intelligence-enabled event adjudication and hematologic predictors. Circ Cardiovasc Qual Outcomes. 2022;15(6):e008007. https://doi.org/10.1161/CIRCOUTCOMES.121.008007.

Hadanny A, Shouval R, Wu J, Gale CP, Unger R, Zahger D, et al. Machine learning-based prediction of 1-year mortality for acute coronary syndrome. J Cardiol. 2022;79(3):342-51. https://doi.org/10.1016/j.jjcc.2021.09.010.

Unterhuber M, Kresoja KP, Rommel KP, Besler C, Baragetti A, Klöting N, et al. Proteomics-enabled deep learning machine algorithms can enhance prediction of mortality. J Am Coll Cardiol. 2021;78(16):1621-31. https://doi.org/10.1016/j.jacc.2021.08.017.

Huang X, Cao T, Chen L, Li J, Tan Z, Xu B, et al. Novel insights on establishing machine learning-based stroke prediction models among hypertensive adults. Front Cardiovasc Med. 2022;9:901240. https://doi.org/10.3389/fcvm.2022.901240.

Van de Leur RR, Bleijendaal H, Taha K, Mast T, Gho JM, Linschoten M, et al. Electrocardiogram-based mortality prediction in patients with COVID-19 using machine learning. Neth Heart J. 2022;30(6):312-8. https://doi.org/10.1007/s12471-022-01665-3.

Kwon JM, Kim KH, Jeon KH, Kim HM, Kim MJ, Lim SM, et al. Development and validation of deep-learning algorithm for electrocardiography-based heart failure identification. Korean Circ J. 2019;49(7):629-39. https://doi.org/10.4070/kcj.2018.0446.

Steele AJ, Denaxas SC, Shah AD, Hemingway H, Luscombe NM. Machine learning models in electronic health records can outperform conventional survival models for predicting patient mortality in coronary artery disease. PLoS One. 2018;13(8):e0202344. https://doi.org/10.1371/journal.pone.0202344.

Ward A, Sarraju A, Chung S, Li J, Harrington R, Heidenreich P, et al. Machine learning and atherosclerotic cardiovascular disease risk prediction in a multi-ethnic population. NPJ Digit Med. 2020;3(1):125. https://doi.org/10.1038/s41746-020-00331-1.

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. https://doi.org/10.1371/journal.pone.0174944.

Dimopoulos AC, Nikolaidou M, Caballero FF, Engchuan W, Sanchez-Niubo A, Arndt H, et al. Machine learning methodologies versus cardiovascular risk scores, in predicting disease risk. BMC Med Res Methodol. 2018;18(1):179. https://doi.org/10.1186/s12874-018-0644-1.

Segar MW, Jaeger BC, Patel KV, Nambi V, Ndumele CE, Correa A, et al. Development and validation of machine learning-based race-specific models to predict 10-year risk of heart failure: a multicohort analysis. Circulation. 2021;143(24):2370-83. https://doi.org/10.1161/CIRCULATIONAHA.120.053134.

Cano J, Fácila L, Gracia-Baena JM, Zangróniz R, Alcaraz R, Rieta JJ. The relevance of calibration in machine learning-based hypertension risk assessment combining photoplethysmography and electrocardiography. Biosensors (Basel). 2022;12(5):289. https://doi.org/10.3390/bios12050289.

Cheung CY, Xu D, Cheng CY, Sabanayagam C, Tham YC, Yu M, et al. A deep-learning system for the assessment of cardiovascular disease risk via the measurement of retinal-vessel calibre. Nat Biomed Eng. 2021;5(6):498-508. https://doi.org/10.1038/s41551-021-00711-5.

Sarraju A, Ward A, Chung S, Li J, Scheinker D, Rodríguez F. Machine learning approaches improve risk stratification for secondary cardiovascular disease prevention in multiethnic patients. Open Heart. 2021;8(2):e001802. https://doi.org/10.1136/openhrt-2021-001802.

Drożdż K, Nabrdalik K, Kwiendacz H, Hendel M, Olejarz A, Tomasik A, et al. Risk factors for cardiovascular disease in patients with metabolic-associated fatty liver disease: a machine learning approach. Cardiovasc Diabetol. 2022;21(1):240. https://doi.org/10.1186/s12933-022-01688-7.

Jamthikar AD, Gupta D, Mantella LE, Saba L, Laird JR, Johri AM, et al. Multiclass machine learning vs. conventional calculators for stroke/CVD risk assessment using carotid plaque predictors with coronary angiography scores as gold standard: A 500 participants study. Int J Cardiovasc Imaging. 2021;37(4):1171-87. https://doi.org/10.1007/s10554-020-02072-9.

Pal M, Parija S, Panda G, Dhama K, Mohapatra RK. Risk prediction of cardiovascular disease using machine learning classifiers. Open Med (Wars). 2022;17(1):1100-13. https://doi.org/10.1515/med-2022-0504.

Sanchez-Cabo F, Rossello X, Fuster V, Benito F, Manzano JP, Silla JC, et al. Machine learning improves cardiovascular risk definition for young, asymptomatic individuals. J Am Coll

Cardiol. 2020;76(14):1674-85. https://doi.org/10.1016/j.jacc.2020.07.061.

Rim TH, Lee CJ, Tham YC, Cheung N, Yu M, Lee G, et al. Deep-learning-based cardiovascular risk stratification using coronary artery calcium scores predicted from retinal photographs. Lancet Digit Health. 2021;3(5):e306-e316. https://doi.org/10.1016/S2589-7500(21)00043-1.

Matheson MB, Kato Y, Baba S, Cox C, Lima JA, Ambale-Venkatesh B. Cardiovascular risk prediction using machine learning in a large Japanese cohort. Circ Rep. 2022;4(12):595-603. https://doi.org/10.1253/circrep.CR-22-0091.

Desai RJ, Wang SV, Vaduganathan M, Evers T, Schneeweiss S. Comparison of machine learning methods with traditional models for use of administrative claims with electronic medical records to predict heart failure outcomes. JAMA Netw Open. 2020;3(1):e1918962. https://doi.org/10.1001/jamanetworkopen.2019.18962.