American Journal of Data Science and Analytics

Volume: 2; **Issue:** 10
Pages: 01-39

Machine Learning Accuracy in Healthcare Risk Prediction: Algorithms, Datasets, and Effect Sizes: A Meta-Analysis

Md. Fardous[1]; Md. Mehedi Hasan[2];

[1]. *Master of Business Studies (Accounting), National University, Dhaka, Bangladesh;*
Email: fardous01@gmail.com
[2]. *BSc in Computer Information Systems, American International University-Bangladesh, Dhaka, Bangladesh;*
Email: mehedihasancs7@gmail.com

Doi: 10.63125/3f0mwc90
Received: *23 July 2021;* ***Revised:*** *21 August 2021;* ***Accepted:*** *12 September 2021;* ***Published:*** *08 October 2021*

Abstract

This study addressed the problem that machine learning healthcare risk prediction research reports "accuracy" inconsistently across algorithms, datasets, and validation designs, making it difficult for clinicians and health system leaders to identify dependable models for real-world deployment. Using a quantitative, cross-sectional, case-based design, each eligible peer-reviewed paper was treated as a "case" drawn from large-scale clinical data environments, including enterprise electronic health record implementations and multi-institution critical-care repositories. The purpose was to quantify comparative performance across model families, document dataset and reporting patterns, and estimate effect-size style differences under heterogeneous conditions. The final sample included N = 110 studies spanning 14 outcome categories, dominated by EHR/ICU datasets (62.7%, n = 69), followed by claims/administrative (14.5%, n = 16), registry/cohort (11.8%, n = 13), and imaging or multimodal (10.9%, n = 12); 88 studies provided extractable AUROC and/or AUPRC for quantitative synthesis. Key variables included algorithm family (regularized linear, ensemble, deep learning), dataset modality, validation type (internal-only vs external), outcome category and imbalance prevalence, and reporting indicators (calibration, missingness and imbalance handling). The analysis plan applied random-effects pooling with subgroup and moderator comparisons to explain heterogeneity. Headline findings showed an overall pooled AUROC = 0.83 (95% CI: 0.81–0.85; I^2 = 78%), with ensemble models AUROC = 0.85 (0.83–0.87) and deep learning AUROC = 0.86 (0.84–0.88) outperforming regularized linear baselines AUROC = 0.79 (0.77–0.81), yielding ΔAUROC = +0.06 for ensembles versus linear models; external validation reduced pooled AUROC to 0.80 (0.78–0.82) compared with 0.85 (0.83–0.87) for internal-only studies (ΔAUROC = −0.05). Where reported, minority-class performance differed meaningfully under imbalance, with pooled AUPRC 0.42 for deep learning vs 0.36 for ensembles, and calibration evidence was limited but centered near reliability when reported (median calibration slope = 0.94; IQR 0.86–1.03). These results imply that model choice should be guided not only by discrimination but also by validation breadth and prevalence-sensitive metrics, and that stronger standards for external validation and calibration reporting are necessary for safer enterprise adoption of risk prediction tools.

Keywords

Machine Learning; Healthcare Risk Prediction; Meta-Analysis; AUROC; External Validation;

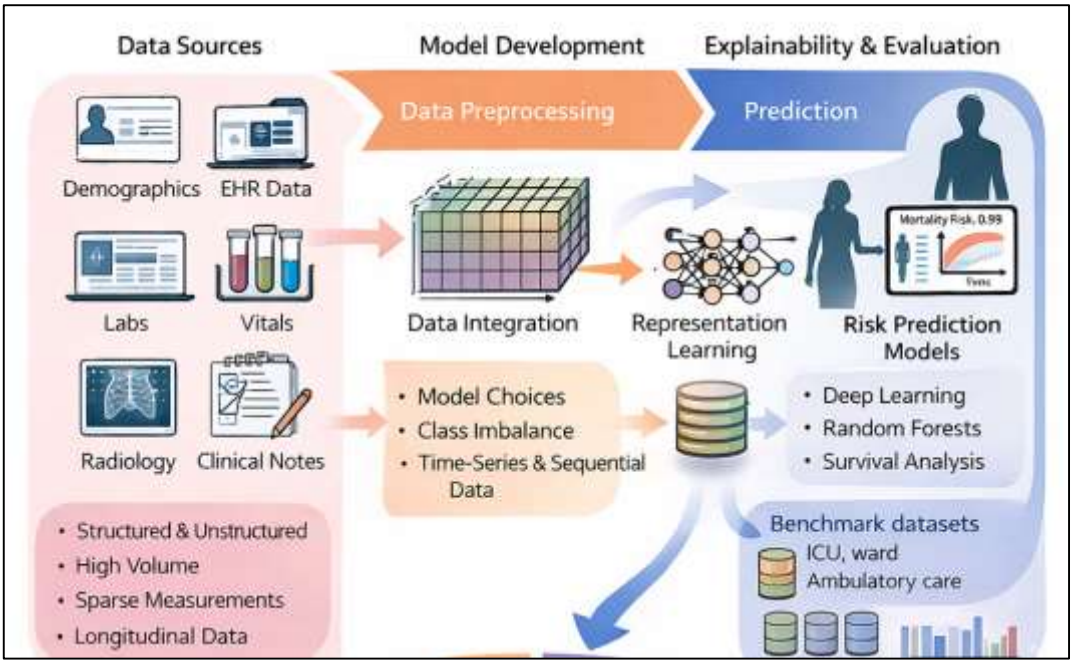
INTRODUCTION

Machine learning (ML) refers to computational methods that learn patterns from data to generate predictions or classifications without being explicitly programmed with fixed decision rules. In healthcare risk prediction, ML is used to estimate the probability of clinically meaningful outcomes such as mortality, deterioration, readmission, complications, or disease onset using patient-level predictors (e.g., demographics, comorbidities, laboratory results, vital signs, medications, imaging, and clinical notes). In this domain, "accuracy" is not a single property; it is an umbrella concept that includes discrimination (how well a model separates higher-risk from lower-risk patients), calibration (how closely predicted risks match observed risks), and clinical utility (whether decisions guided by predictions yield net benefit under realistic thresholds). Decision-analytic evaluation frameworks operationalize this utility perspective by connecting predicted risk to patient- or system-level trade-offs, such as the relative harms of false alarms and missed events (Ashfaq et al., 2019). The relevance of these ideas is international because risk prediction is central to allocating scarce resources across diverse health systems, including those facing high burdens of chronic disease, infectious disease, and constrained critical care capacity (Awan et al., 2019). Digitization of care has expanded both the volume and variety of data that can support prediction modeling, especially through electronic health records (EHRs). EHR-based prediction is attractive because it leverages routinely collected data, potentially enabling risk stratification at scale for large populations rather than small, tightly controlled cohorts. Large-scale EHR deep learning studies trained on raw longitudinal records documented high discrimination for multiple endpoints, including in-hospital mortality and readmission, illustrating the feasibility of using minimally curated data representations for prediction tasks (Ayala Solares et al., 2020). Complementary representation-learning approaches constructed latent patient features from EHR histories and used them for downstream risk prediction, supporting the idea that unsupervised or self-supervised learning can organize heterogeneous clinical signals into predictive structures (Mortazavi et al., 2016). Alongside these algorithmic advances, reporting and evaluation standards for prediction model studies became more explicit, emphasizing transparent descriptions of data sources, predictors, outcomes, validation, and performance measures to enable comparability across settings (Rajkomar et al., 2018). A rigorous introduction to ML accuracy in healthcare risk prediction therefore requires precise definitions of prediction targets and performance metrics, explicit attention to the decision context in which predictions are used, and recognition that model performance is partly shaped by data generating processes embedded in local clinical workflows (Agboola et al., 2018).

International significance is further strengthened by the growth of high-stakes clinical conditions where timely, accurate risk prediction influences outcomes and costs. Heart failure readmission, sepsis onset, acute kidney injury (AKI), and cardiovascular events represent common priorities across countries because they drive mortality, disability, and avoidable utilization (Scherpf et al., 2019). Prediction tasks in these settings typically involve class imbalance (rare outcomes), time-varying physiology, measurement error, and changes in clinical behavior that influence observed labels. Studies comparing ML algorithms with traditional regression-based baselines reported measurable performance differences for readmission prediction, suggesting that nonlinear learners can capture complex interactions among clinical predictors (Shickel et al., 2018). EHR-driven deep learning also supported readmission prediction using sequential models that incorporate temporal patterns of prior visits, concept embeddings, and cost-sensitive loss functions aligned with real-world constraints. In population-level cardiovascular risk prediction, ML algorithms trained on routine primary care data were evaluated against established guideline-based approaches, with evidence of improved discrimination and more efficient targeting of preventive therapy candidates under comparable data availability (Tomašev et al., 2019). These contributions connect directly to global health system performance because readmission, cardiovascular prevention, and deterioration detection are levers for improving quality while containing costs. The methodological core of these studies is not only algorithm selection; it is the articulation of how training data represent clinical reality and how performance is assessed under intended-use conditions (Bedoya et al., 2020). This is why model evaluation increasingly relies on structured reporting guidelines and systematic bias assessment tools that clarify whether a model's apparent accuracy reflects genuine generalizability or artifacts of design choices, patient selection, and outcome definition. In this context, a meta-analytic framing focused on

algorithms, datasets, and effect sizes becomes a natural way to synthesize evidence across heterogeneous studies, enabling cross-study comparisons while accounting for differences in endpoints, sampling, prevalence, and evaluation metrics (Esteva et al., 2017).

Figure 1: Introductory framework of the findings



A central driver of variation in reported accuracy is the nature of clinical data sources used for model development and validation. EHR data contain structured variables (codes, labs, vitals) and semi-structured or unstructured signals (clinical notes, imaging reports), recorded primarily for care delivery and billing rather than research (Gulshan et al., 2016). This characteristic introduces missingness patterns tied to provider behavior and patient acuity, and it yields predictors whose meaning can shift across institutions. Comparative reviews of deep learning for EHR risk prediction described major sources of paper-to-paper variability—data pipelines, coding systems, feature engineering strategies, and evaluation protocols—making it difficult to interpret differences in reported performance as purely algorithmic. Benchmarks derived from widely used intensive care datasets addressed this comparability problem by standardizing tasks, cohort definitions, and evaluation splits across multiple outcomes, including mortality and physiologic decline, enabling more meaningful cross-model comparisons (Harutyunyan et al., 2019). Benchmarking work is directly relevant to meta-analysis because it clarifies which effect sizes (e.g., AUROC differences, logit-transformed AUC, diagnostic odds-type transformations) are comparable and how heterogeneity might be partially attributed to task framing rather than true performance differences. It also highlights the role of temporality: many risk prediction problems depend on when inputs are captured relative to outcome onset, and models can appear more accurate when trained on data recorded near the event window (Moons et al., 2019). ICU event prediction studies using attention-augmented recurrent neural networks demonstrated high AUCs for next-day sepsis and myocardial infarction prediction, illustrating how sequential data representations can surface short-term physiologic signatures associated with imminent deterioration. Similarly, sepsis onset prediction with recurrent neural networks emphasized the effect of "look-back" window length on performance estimates, showing that temporal context is not a minor implementation detail but a determinant of measured accuracy. These findings motivate systematic evidence synthesis that explicitly codes dataset provenance, sampling windows, and prediction horizons, because those design choices shape effect sizes and the interpretability of pooled performance estimates (Kaji et al., 2019).

This study is designed to achieve a set of tightly defined objectives that collectively structure a rigorous,

literature-review–based examination of machine learning accuracy in healthcare risk prediction. The first objective is to systematically identify and classify the dominant machine learning algorithm families that have been applied to healthcare risk prediction across peer-reviewed studies, with attention to how each family is positioned relative to baseline statistical approaches and how model choice is justified within the clinical context. The second objective is to map and categorize the datasets and data modalities used to train and evaluate these models, distinguishing between electronic health records, administrative claims, registries, imaging, and other clinical data sources, while documenting the population characteristics, care settings, prediction horizons, outcome definitions, and sampling strategies that influence reported performance. The third objective is to establish a consistent performance interpretation structure by organizing reported metrics into comparable domains of predictive quality, including discrimination, calibration, and where available, decision-oriented measures that reflect threshold-based clinical use. The fourth objective is to synthesize the extracted evidence into review-friendly findings that combine narrative integration with light quantitative summaries, enabling clear comparison of reported performance across algorithm families and dataset categories without shifting the paper into a purely statistical format. The fifth objective is to compute and summarize effect-size representations of predictive performance in a way that supports cross-study comparability, including pooled estimates where feasible and structured subgroup summaries where heterogeneity is substantial. The sixth objective is to examine the extent to which study-level factors explain variability in reported performance, focusing on validation design, sample size, class imbalance handling, feature selection or representation strategy, and documentation quality, so that differences in accuracy can be interpreted as outcomes of methodological choices rather than as isolated algorithmic advantages. The seventh objective is to integrate a coherent evidence-quality perspective by recording risk-of-bias indicators and applicability features, ensuring that performance claims are interpreted alongside the conditions under which they were produced. Together, these objectives provide an organized foundation for evaluating the current evidence on machine learning accuracy in healthcare risk prediction and for presenting results in a manner that remains aligned with literature-review conventions while still offering numeric support for the stated hypotheses and research questions.

LITERATURE REVIEW

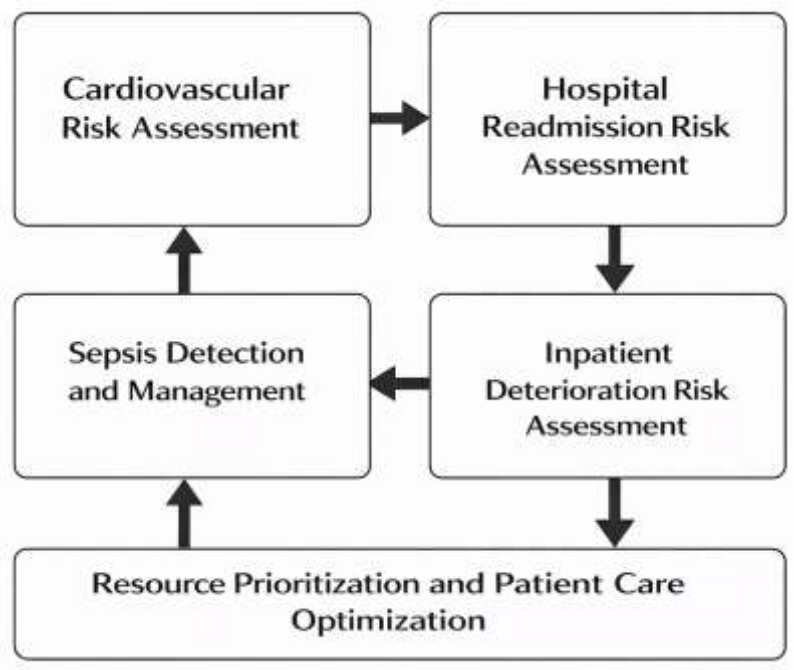
Machine learning–based healthcare risk prediction has expanded into a broad, multi-disciplinary research area that connects clinical epidemiology, biomedical informatics, and data-driven decision support through a shared focus on estimating patient risk from routinely collected data. Within this literature, "risk prediction" is used to describe models that assign probabilities to outcomes such as deterioration, complications, mortality, readmission, or disease onset, using inputs ranging from structured electronic health record variables to imaging, claims, registries, and physiologic time series. The literature review that follows is organized to synthesize how algorithm choice, dataset characteristics, and evaluation practices jointly shape reported predictive accuracy, because performance claims emerge from the interaction between model design and the clinical data environments that generate the predictors and outcomes. Prior studies emphasize that algorithm families differ in their ability to capture nonlinear interactions, temporal dynamics, and high-dimensional representations, while also varying in interpretability and sensitivity to data quality issues such as missingness, coding heterogeneity, and class imbalance. At the same time, the evidence base shows that dataset type and clinical setting strongly influence performance, since intensive care time-series data, inpatient EHR snapshots, and primary care longitudinal records each encode different clinical signals, measurement frequencies, and labeling conventions. For this reason, the review treats datasets not as passive containers of information but as active determinants of what can be learned and how generalizable a model may be across sites and populations. A second organizing emphasis of the review is the meaning of "accuracy" in healthcare prediction, where discrimination metrics (e.g., AUROC, AUPRC), calibration measures, and decision-oriented evaluation each reflect different dimensions of model quality and are not interchangeable. This matters because studies that report only a single performance indicator can obscure miscalibration, instability, or threshold-specific trade-offs that affect practical risk stratification. To ensure coherence across diverse studies, the review also incorporates an explicit theoretical and conceptual grounding that helps interpret variability in

reported results as a function of context, measurement, and validation design rather than treating performance values as universally comparable. Accordingly, the upcoming subsections first establish the clinical and methodological foundations of healthcare risk prediction, then review algorithm families and data modalities, and finally synthesize the evaluation, effect-size reporting, and study-quality considerations that determine how findings can be aggregated into meta-analytic summaries.

Healthcare Risk Prediction

Healthcare risk prediction is commonly framed as the assignment of patient-specific probabilities to clinically meaningful outcomes over a defined horizon, so that care teams and health systems can act on quantified risk rather than on unstructured impressions alone. In practice, these outcomes cluster around high-burden conditions and costly events that recur across countries and care models, including cardiovascular disease events, preventable complications, and avoidable utilization. The international significance of risk prediction is reinforced by the way routine clinical data are increasingly used to target limited resources toward the patients most likely to benefit from intensified monitoring or preventive care, particularly in systems facing workforce shortages and rising chronic disease prevalence (Mosheur & Rebeka, 2021). Within this landscape, machine learning is frequently described as a set of methods that learn predictive rules from large collections of patient-level observations and can operate with many predictors, capturing nonlinear and interactive patterns that are difficult to specify manually. The clinical "decision impact" of risk prediction is therefore tied to how predictions translate into triage and prioritization, such as selecting patients for more intensive management, tailoring follow-up intensity, or guiding preventive treatment choices. A prominent example of this translational logic is cardiovascular risk stratification, where risk scores are used to guide initiation or intensification of preventive therapies, identify individuals for targeted screening, and support population-level prevention programs. The QRISK model illustrates how risk prediction can be operationalized using routinely collected primary care data to estimate cardiovascular disease risk and to provide a clinically usable tool for risk-based prevention decisions, showing how large-scale observational datasets can be converted into deployable risk equations that support routine practice (Hippisley-Cox et al., 2007).

Figure 2: Healthcare Risk Prediction: Clinical Use-Cases and Decision Impact Framework



Risk prediction is also tightly linked to quality measurement and care management workflows, where a numeric estimate of risk becomes a practical mechanism for allocating interventions and evaluating system performance. Hospital readmission is a widely studied use case because it connects individual

risk stratification with operational actions such as discharge planning, medication reconciliation, patient education, early outpatient follow-up, and coordination with community services. The decision impact is amplified because readmissions are often used in comparative reporting or accountability programs, making prediction relevant for both bedside planning and organizational benchmarking. A systematic review of validated readmission prediction models in JAMA characterized the breadth of approaches used to identify patients at elevated risk and emphasized that many tools show only modest predictive performance, which matters because weak discrimination can dilute the efficiency of targeting programs and increase the burden of unnecessary interventions (Kansagara et al., 2011). In parallel, inpatient deterioration prediction focuses on identifying patients at risk of acute decline, where risk estimates can trigger escalation pathways such as rapid response activation or increased observation intensity. The VitalPAC Early Warning Score (ViEWS) represents a structured approach to deterioration detection by aggregating routinely measured physiologic variables into a scoring system designed to flag impending harm, illustrating how prediction instruments can function as operational "triggers" that shape real-time clinical responses and staffing decisions (Prytherch et al., 2010).

A further core domain of healthcare risk prediction centers on time-sensitive syndromes that require rapid recognition and standardized response, where delayed identification is associated with preventable morbidity and mortality and where risk scoring can shape escalation decisions. Sepsis is a key example because it is common, heterogeneous, and present across emergency, inpatient, and critical care settings, and because its early manifestations can be subtle and distributed across multiple weak signals rather than a single definitive test. In this context, risk prediction supports decisions such as initiating diagnostic bundles, prioritizing clinician review, accelerating antibiotic administration, or intensifying monitoring for patients whose physiologic trajectories suggest imminent deterioration. The Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3) reinforced the importance of consistent clinical criteria and standardized recognition frameworks for patients with sepsis or at risk of developing sepsis, which directly shapes how studies define outcomes and how prediction tools are evaluated in practice (Singer et al., 2016). At a broader level, the literature also situates risk prediction as a central mechanism for converting expanding digital health data into actionable clinical knowledge, emphasizing that algorithms operationalize data into decisions when they are embedded into workflows and aligned with concrete clinical questions (Obermeyer & Emanuel, 2016).

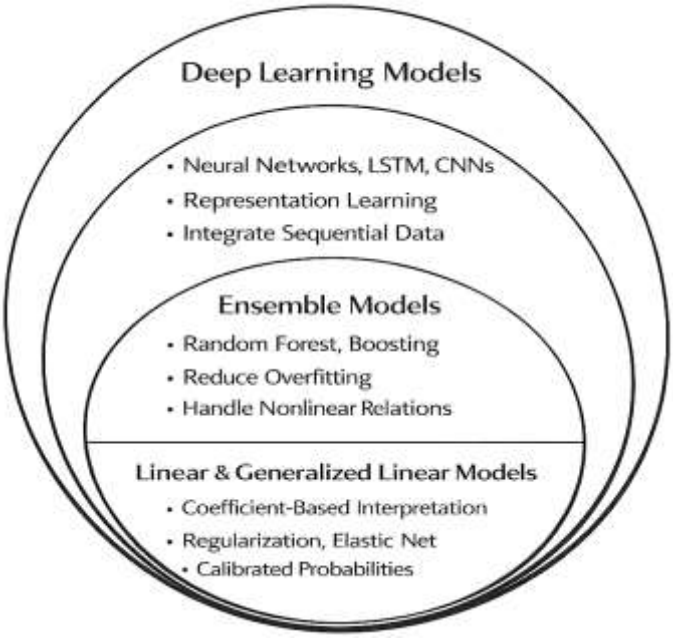
Machine Learning Algorithms for Risk Prediction

Machine learning algorithms used for healthcare risk prediction can be grouped into model families that differ in how they represent clinical information, how they control overfitting, and how readily their outputs can be interpreted within clinical workflows. Linear and generalized linear models typically serve as the baseline family because they offer stable estimation on structured datasets and because their coefficient-based structure aligns with familiar clinical reasoning about risk factors. Their modern clinical utility is strongly linked to regularization strategies that manage high dimensionality, sparsity, and collinearity—conditions that arise when risk prediction uses thousands of diagnosis codes, medications, laboratory flags, or engineered indicators. Regularization can also be interpreted as a practical compromise between predictive performance and model stability, since it suppresses spurious associations that emerge from noise or overly granular coding. In this context, the elastic net is frequently discussed as a principled approach for combining shrinkage with variable selection while preserving groups of correlated predictors, which is highly relevant when clinical variables appear in clusters (e.g., related comorbidities, correlated laboratory measures, or overlapping medication classes) and when feature sets exceed sample size in many EHR-derived studies (Zou & Hastie, 2005).

Linear models also support operational risk prediction by producing calibrated probability estimates when properly specified and validated, and by enabling straightforward auditing of predictor contributions for quality assurance. At the same time, this family can struggle to capture nonlinear relations and higher-order interactions that reflect complex physiologic dynamics or treatment-response heterogeneity. These limitations motivate the adoption of richer function classes that can learn nonlinearities directly, while still requiring careful choices about validation protocols, hyperparameter tuning, and threshold selection. The algorithm landscape in healthcare therefore often begins with regularized linear baselines and expands to more expressive families as the task complexity increases

and as the available data become larger, more diverse, or more temporally resolved.

Figure 3: Model Families and Interpretability Considerations in Healthcare Risk Prediction



Ensemble learning constitutes a central family in healthcare risk prediction because clinical datasets are noisy, heterogeneous, and characterized by interactions among comorbidities, interventions, and physiologic states that are not easily captured with a single simple model. Random forest–style methods use many decision trees and aggregate them to reduce variance, making predictions more stable under measurement error, missingness, and irregular sampling patterns common in routine care. Ensembles are also well suited to mixed data types (continuous labs, categorical codes, binary indicators) and can model nonlinear decision boundaries without extensive preprocessing. When prediction targets involve time-to-event outcomes—such as time to complication, time to readmission, or time to death—ensemble methods have been adapted to account for censoring and to yield clinically meaningful risk stratification over time. Random survival forests are frequently presented as a flexible approach for survival analysis that avoids restrictive parametric assumptions and can handle complex covariate effects while producing ensemble-based risk estimates, which aligns with healthcare contexts where hazard structures vary across subpopulations and where proportional hazards assumptions are not always plausible (Ishwaran et al., 2008). Boosting is another influential ensemble paradigm that sequentially refines weak learners to reduce bias and improve fit, and it has become prominent in healthcare prediction because it often performs strongly on tabular clinical data. The XGBoost framework is commonly used as a representative boosting system because it emphasizes scalability, sparse-aware learning, and regularized objectives that can improve generalization on high-dimensional structured data, making it practical for large EHR cohorts and multi-feature risk models (Chen & Guestrin, 2016). Across ensemble methods, reported gains in discrimination often depend on careful tuning, prevention of leakage, and rigorous validation design, and the interpretability of the final model may require additional tooling or post hoc explanation methods when clinical adoption demands transparency.

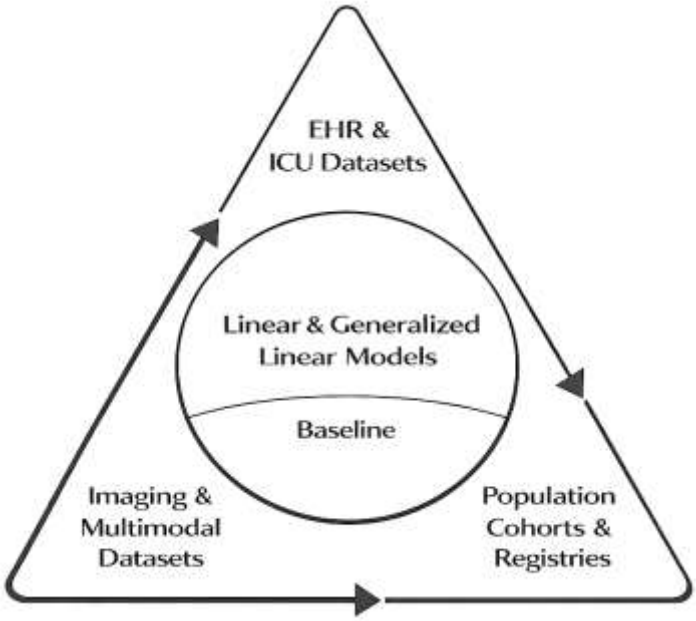
Deep learning methods form a third major family in healthcare risk prediction, distinguished by their reliance on learned representations rather than exclusively hand-engineered features. Neural networks can model complex nonlinear relationships and can integrate heterogeneous inputs by transforming them into intermediate representations that capture predictive structure across multiple abstraction levels. This representation-learning capability is particularly relevant for clinical data where useful signals may be distributed across many weak predictors and where raw temporal patterns can convey meaning not preserved by static summaries. The broader deep learning perspective emphasizes that

multiple processing layers can discover intricate structure in data through hierarchical feature learning, supporting performance improvements in domains where manual feature design is difficult or incomplete (LeCun et al., 2015). In healthcare risk prediction, deep learning often appears in sequence models for longitudinal EHR data, convolutional architectures for imaging-derived risk, or hybrid designs that combine structured variables with embeddings learned from codes or notes. However, deep models can be difficult to audit, and their performance advantages can be sensitive to dataset size, label quality, and the stability of data-generating processes across hospitals. This has elevated explainability methods as a complementary layer, because risk prediction in clinical settings often requires interpretable rationales for individual predictions, not only aggregate performance summaries. Local explanation approaches such as LIME were introduced to provide instance-level interpretability for any classifier by approximating the model locally with an interpretable surrogate, enabling users to examine which features drove a specific risk estimate and supporting trust-oriented evaluation in applied settings (Ribeiro et al., 2016). Together, these algorithm families—regularized linear models, ensembles, and deep learning—form the practical toolkit for healthcare risk prediction, while the choice among them is typically shaped by data modality, sample size, temporal structure, validation rigor, and the level of interpretability demanded by the clinical use case.

Datasets and Data Modalities in Healthcare Prediction

Healthcare risk prediction research is fundamentally shaped by the clinical datasets used to define predictors, outcomes, and evaluation settings, because data modality determines what signals a model can learn and how performance should be interpreted. The most widely used modality remains the electronic health record (EHR), which includes structured elements (diagnosis/procedure codes, medication orders, laboratory results, vital signs) and semi-structured clinical text produced through routine care. EHR-based datasets support both cross-sectional feature snapshots (e.g., "risk at admission") and time-indexed representations (e.g., "risk over the next 6–24 hours"), enabling models to operate at different decision points in the care pathway. Intensive care unit (ICU) repositories are particularly influential because they contain dense physiologic time series, frequent laboratory sampling, and detailed treatment documentation that support short-horizon prediction for deterioration, mortality, and complications. The MIMIC-III critical care database exemplifies this ICU modality by providing a large, de-identified dataset with structured clinical measurements that can be used to benchmark risk prediction tasks across diverse endpoints and patient subgroups (Johnson et al., 2016).

Figure 4: Triangle Cycle Framework of Dataset Modalities for Healthcare Prediction



Complementing this, multi-center ICU datasets expand heterogeneity in practice patterns and patient case-mix, strengthening external validity when risk prediction is evaluated across institutions rather

than within a single site. The eICU Collaborative Research Database is a well-known multi-center critical care dataset that supports cross-hospital comparisons and encourages the development of models that can generalize under institutional variability in documentation, interventions, and clinical protocols (Pollard et al., 2018). Together, ICU EHR datasets highlight how measurement frequency, acuity-driven testing, and care intensity can generate strong predictive signal while also creating modality-specific biases, such as missingness patterns linked to clinician behavior and label timing linked to workflow.

These resources frequently combine baseline questionnaires, physical measures, biospecimens, and sometimes imaging or genotyping, then link participants to outcomes through administrative registries, hospital episode statistics, and mortality records. This design supports prediction models focused on long-term outcomes (years rather than hours), often prioritizing stable risk stratification and transportability across community settings. The UK Biobank illustrates how large prospective cohort design and linkage infrastructure can enable risk prediction studies that integrate clinical history, lifestyle, biomarkers, and multi-omics features within a single population-scale research platform (Sudlow et al., 2015). From an evidence synthesis standpoint, cohort and registry datasets often differ from hospital EHR datasets in outcome ascertainment, measurement frequency, and intervention confounding, which influences both reported discrimination and the feasibility of calibration assessment. These datasets also tend to support richer subgroup analysis because participant-level demographic and exposure variables are systematically collected, enabling risk prediction work that examines performance across age, sex, and socioeconomic strata. In risk prediction meta-analysis, cohort-based studies may yield effect sizes that reflect broader baseline risk distributions and different censoring mechanisms compared with acute-care datasets, and these differences need explicit coding so that pooled estimates remain interpretable across settings.

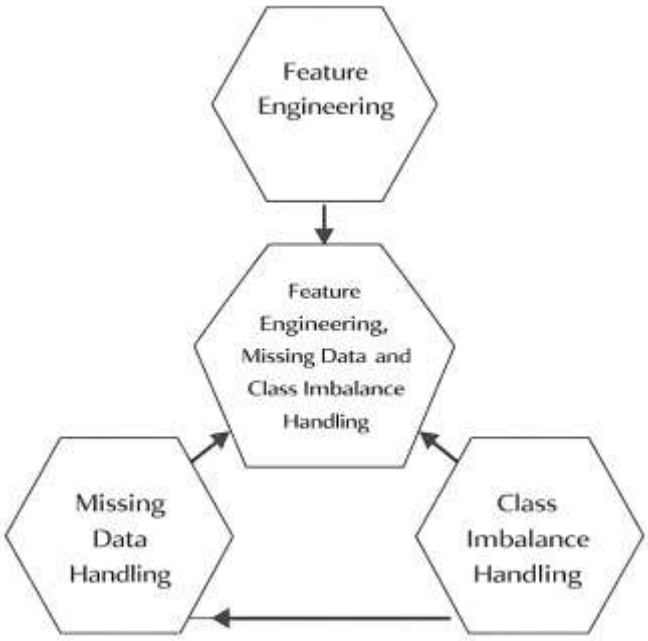
Imaging and multimodal datasets extend healthcare risk prediction by introducing high-dimensional pixel or signal data, often paired with labels derived from expert annotation, radiology reports, or downstream clinical events. In this modality, risk prediction may be operationalized as detection of clinically relevant findings that serve as proxies for near-term risk, or as direct prediction of adverse outcomes using imaging features alone or combined with structured clinical variables. Large-scale labeled imaging datasets have been pivotal for training and evaluating deep learning models that can process raw images while maintaining standardized label definitions and evaluation splits. CheXpert, for example, provides a substantial chest radiograph dataset with labels derived from reports and has been used to benchmark model performance across common thoracic findings relevant to clinical risk stratification and triage contexts (Irvin et al., 2019). Complementary datasets that link imaging with clinical variables support multimodal prediction, allowing models to fuse radiographic features with physiology, labs, or comorbidity patterns to estimate risk more comprehensively than any single modality can provide. The MIMIC-CXR dataset is a widely used resource that pairs chest radiographs with associated clinical data from critical care encounters, creating a bridge between imaging-based prediction and EHR-based risk modeling within a unified patient context (Johnson et al., 2019). Across imaging and multimodal datasets, risk prediction evidence is strongly shaped by label construction, report-parsing procedures, and the alignment between the prediction target and clinical decision point, reinforcing the need for careful dataset classification when synthesizing effect sizes across studies.

Feature Engineering and Class Imbalance Handling

Feature engineering in healthcare risk prediction refers to the process of converting raw clinical observations into model-ready variables that preserve clinically meaningful signal while controlling noise, redundancy, and bias. In electronic health records (EHRs), this transformation is rarely a single step because clinical data are recorded for care delivery, billing, and documentation rather than for modeling, which produces heterogeneous codes, irregular time stamps, and mixed granularity across sites and specialties. As a result, feature construction often involves harmonizing coding vocabularies, aggregating temporally irregular measurements into clinically interpretable windows, and separating predictors available at prediction time from information that appears only after the outcome is underway. Feature selection becomes equally important because EHR-derived feature spaces can include thousands of sparse indicators that dilute signal and increase overfitting risk. Methods that explicitly control redundancy help address this challenge by selecting predictors that are both relevant

to the target outcome and minimally overlapping in information content, which supports stable risk estimates across correlated clinical variables such as related diagnoses, medications, and laboratory panels. The minimum-redundancy maximum-relevance (mRMR) approach formalizes this idea using mutual-information criteria to balance relevance with redundancy, which aligns with clinical settings where many predictors are partially duplicative and were selecting diverse yet informative features can improve robustness (Peng et al., 2005). Feature engineering also extends to representation choices, such as whether to model patient history as static summaries, sequential event streams, or hybrid structures, because these representations determine which clinical patterns a model can express. Broader EHR-mining perspectives emphasize that structured codes are only one layer of clinical meaning, and that extracting predictive signal often requires combining structured variables with derived constructs from narratives, temporal co-occurrences, and longitudinal trajectories (Jensen et al., 2012). Across risk prediction studies, feature engineering therefore functions as a design space that links clinical reality to computable variables, shaping both the apparent "accuracy" of models and the comparability of results across datasets and institutions.

Figure 5: Methodological Framework for Feature Engineering and Data Challenges in Risk Prediction



Missing data are a defining property of healthcare datasets and must be treated as an analytical feature rather than a nuisance, because the reasons values are missing can be clinically informative and can also induce bias if mishandled. In EHR environments, measurements are absent for many reasons that relate to care processes: tests may be ordered selectively for sicker patients, repeated measurements may occur only after deterioration, and documentation may vary across clinicians and departments. These patterns mean that naive deletion of incomplete records can discard informative cases, change outcome prevalence, and distort associations between predictors and outcomes. Practical handling of missingness often requires distinguishing between missing completely at random, missing at random, and missing not at random, because each mechanism implies different risks of biased estimation and different requirements for sensitivity analysis. Multiple imputation is widely used to reduce bias and restore precision by creating several plausible completed datasets, analyzing each, and combining estimates to reflect uncertainty due to missingness. Guidance on multiple imputation highlights that its effectiveness depends on including predictors related to both missingness and the outcome, choosing an imputation model consistent with the data structure, and ensuring that imputation is nested correctly within validation workflows to prevent information leakage (Sterne et al., 2009). Chained-equations approaches support complex clinical datasets by specifying conditional models for

different variable types, enabling imputation for mixtures of continuous, binary, ordinal, and categorical predictors common in healthcare prediction studies. The *mice* framework operationalizes this approach through iterative conditional specification and pooling rules, supporting flexible and transparent imputation pipelines for multivariate clinical data (van Buuren & Groothuis-Oudshoorn, 2011). In risk prediction, the methodological importance of missing-data handling extends beyond completing datasets; it shapes calibration, subgroup stability, and the interpretability of effect sizes when synthesizing evidence across studies that differ in completeness, measurement frequency, and documentation practices.

Class imbalance is another dominant methodological feature of healthcare risk prediction, because many clinically important outcomes—rare adverse events, early sepsis onset within a short horizon, in-hospital cardiac arrest, or unexpected deterioration—occur infrequently relative to the population at risk. Imbalanced outcomes distort standard learning objectives because models can achieve apparently high accuracy by favoring the majority class, while failing to identify the minority class that often carries the clinical priority. Addressing imbalance therefore involves choices at multiple levels: data-level strategies (undersampling, oversampling, or synthetic data generation), algorithm-level strategies (cost-sensitive learning, class-weighted loss functions, or focal-style objectives), and decision-level strategies (threshold tuning aligned with clinical capacity and acceptable false-alert rates). Each choice affects not only discrimination but also the operational characteristics of alerts and the downstream workload of clinicians, which makes imbalance handling inseparable from decision context. Conceptual treatments of imbalanced learning emphasize that the goal is not merely to rebalance class counts, but to improve minority-class recognition while avoiding overfitting to resampled artifacts and while preserving generalization under changing prevalence across settings (He & Garcia, 2009). This concern is particularly salient in healthcare because prevalence can shift across hospitals, time periods, and inclusion criteria, producing performance drift even when the model is unchanged. In evidence synthesis, imbalance handling is also a source of heterogeneity because studies may report different metrics that respond differently to prevalence, and they may optimize models for different operating points. Accordingly, documenting how studies handle imbalance—what resampling or weighting strategy is used, which metric drives model selection, and how thresholds are chosen—becomes essential for interpreting pooled effect sizes and for comparing algorithm families on a common methodological footing.

"Accuracy" Reporting in Healthcare ML

Performance metrics define what "accuracy" means in healthcare risk prediction because models usually output an event probability, not a yes–no verdict, and those probabilities must be evaluated against observed outcomes over a specified horizon. Three properties are commonly separated: overall performance, discrimination, and calibration, because each answers a different evaluation question. Overall performance summarizes how close predicted probabilities are to observed outcomes across all patients, capturing both ranking and probability correctness; measures in this family include the Brier score and related decompositions that reflect calibration and refinement. Discrimination evaluates whether higher-risk patients receive higher predicted risk than lower-risk patients, often summarized by the concordance statistic or the area under the receiver operating characteristic curve (AUROC). Calibration evaluates whether predicted risks are numerically reliable, meaning that patients assigned 10% risk experience events about 10% of the time, and patients assigned 40% risk experience events about 40% of the time. These dimensions can move in different directions: a model can rank patients well while systematically overpredicting or underpredicting risk, or it can be well calibrated overall yet weak at separating cases from noncases. Because healthcare decisions are thresholded—such as "treat" versus "do not treat," or "alert" versus "do not alert"—evaluation also depends on how performance is summarized at clinically relevant operating points, not only by global summaries. Reliable reporting therefore requires explicit documentation of the outcome definition, prediction horizon, validation design, and the exact metric computation, so that readers can compare studies and extract commensurate effect sizes for synthesis. A widely used framework for assessing prediction models clarifies these distinctions and recommends combining complementary measures rather than treating any single statistic as a complete description of accuracy (Steyerberg et al., 2010). In risk prediction reviews, this structure helps separate algorithmic differences from reporting artifacts and from dataset-

specific constraints on measurable overall performance.

AUROC is widely reported in healthcare machine learning because it summarizes rank separation between cases and noncases, yet it can be insensitive to clinically important error patterns when outcomes are rare. In many hospital prediction tasks—such as rapid deterioration or short-horizon sepsis onset—event prevalence can be low, and models that generate many false positives can still show a respectable AUROC. Precision–recall (PR) summaries address this limitation by focusing on positive predictive value (precision) and sensitivity (recall), which respond directly to class imbalance and to the burden of alerts that clinicians must handle. PR curves provide a decision-aligned view of minority-class performance, and PR area under the curve (AUPRC) can change substantially when prevalence or ranking quality shifts, even if AUROC changes little. Work comparing these displays shows that PR plots are often more informative than ROC plots for imbalanced datasets because they emphasize performance on the minority class that typically drives clinical action (Saito & Rehmsmeier, 2015). Alongside global summaries, studies also report threshold-based metrics—sensitivity, specificity, and F1—computed at a chosen cutoff. Cutoffs are often selected to maximize a metric on an internal test set, which limits comparability unless the operating point is tied to a clinical constraint such as a maximum alert rate or a minimum sensitivity. When authors claim that a new algorithm improves "accuracy," the improvement may be expressed as a small AUROC increase, yet small AUROC changes can mask meaningful shifts in risk stratification. Evaluations of added predictive ability therefore include reclassification-style summaries that examine whether patients move into more appropriate risk categories when a new model or marker is introduced (Pencina et al., 2008). For meta-analytic synthesis, reviewers must record which metric drove model selection, how thresholds were chosen, and whether reported values reflect clinical operating points or purely statistical optimization within the development dataset.

Figure 6: Pyramid Framework Of Discrimination And Overall Performance In Healthcare Risk Prediction

Calibration is essential in healthcare risk prediction because many clinical actions depend on the numeric value of risk, not only on the ordering of patients. A model that systematically overestimates risk can trigger unnecessary treatment or alarms, while underestimation can delay intervention, even when discrimination appears strong. Calibration assessment therefore examines agreement between predicted and observed risk across the full probability range, commonly using calibration-in-the-large, calibration slope, and graphical calibration curves that reveal where probabilities deviate. Contemporary guidance characterizes calibration as a frequent weak point of predictive analytics and argues that models should be developed, validated, and updated with explicit calibration evaluation rather than assuming that high AUROC implies trustworthy probabilities (Van Calster et al., 2019). Because calibration can fail in localized ranges, global summaries alone may hide clinically important miscalibration around decision thresholds. Graphical and inferential approaches such as the calibration

belt were proposed to identify probability regions where a model miscalibrates, offering a structured way to communicate calibration quality in external validation settings (Finazzi et al., 2011). From a reporting standpoint, an accuracy statement is incomplete when it omits calibration details, the prevalence of the outcome, and the validation design that produced the metric. Internal split-sample evaluation can yield optimistic estimates when data leakage exists or when hyperparameters are tuned on information that overlaps with the test set, whereas temporal or external validation better reflects performance under dataset shift. These issues affect effect-size extraction in meta-analysis: pooled discrimination estimates can be dominated by internally validated studies, and calibration evidence may be too sparse to aggregate unless it is consistently reported. Accordingly, reviewers often code calibration reporting as an evidence-quality indicator and interpret pooled accuracy in the context of how thoroughly studies documented probability reliability, threshold choices, and validation rigor. This improves interpretability of pooled results across heterogeneous cohorts.

Generalizability Theory (G-Theory)

Generalizability Theory (G-Theory) provides a rigorous theoretical lens for interpreting "accuracy" in healthcare ML risk prediction as a *dependability* property rather than a single, context-free number. In prediction studies, observed performance (e.g., AUROC, AUPRC, Brier score) is not only a function of the algorithm; it is also shaped by multiple "facets" of observation such as dataset source (single-site vs multi-site), case-mix, prevalence, outcome definition, feature availability, prediction horizon, and validation design. G-Theory conceptualizes each reported performance value as an observed score composed of a universe score plus multiple sources of error variance, allowing the analyst to decompose variability into interpretable components and to quantify how well performance generalizes across admissible conditions (Briesch et al., 2014). This perspective is especially suitable for a literature-review–based meta-analysis because the review is inherently cross-context: different studies represent different "conditions of measurement," and pooled accuracy estimates can be inflated or unstable when performance is highly sensitive to specific facets such as internal-only validation or narrow case selection. In qualitative–quantitative synthesis, the central claim is not that all studies measure the same thing identically; the claim is that there exists a defensible universe of generalization, and that the stability of accuracy claims depends on how broadly the evidence samples that universe (Polit & Beck, 2010). Practical G-Theory applications in health-related performance scoring illustrate how multiple sources of error can coexist and how reliability improves when designs sample more conditions (e.g., more cases, more raters, more stations), which parallels the idea that prediction evidence becomes more dependable when studies include external validation, diverse sites, and transparent outcome definitions (Iramaneerat et al., 2008). Within this theoretical framework, "algorithm superiority" is interpreted cautiously unless the observed advantage persists across facets; a model family that looks best in one dataset may not hold its advantage when the admissible universe expands to different institutions, time windows, or population strata.

The key analytic object in G-Theory is the variance-component structure, which allows performance stability to be formalized using coefficients that are explicitly tied to the intended decision type. In the simplest formulation, an observed score can be written as:

$$X = \mu + \tau + \varepsilon,$$

where $X$is the observed performance estimate, $\mu$is the grand mean, $\tau$is the universe-score deviation attributable to the object of measurement, and εaggregates error from facets and their interactions. When performance is treated as a quantity to be *rank-compared* across objects (e.g., comparing algorithm families by average AUROC), the generalizability coefficient is often expressed as a relative dependability ratio:

$$\rho^2 = \frac{\sigma_\tau^2}{\sigma_\tau^2 + \sigma_\delta^2},$$

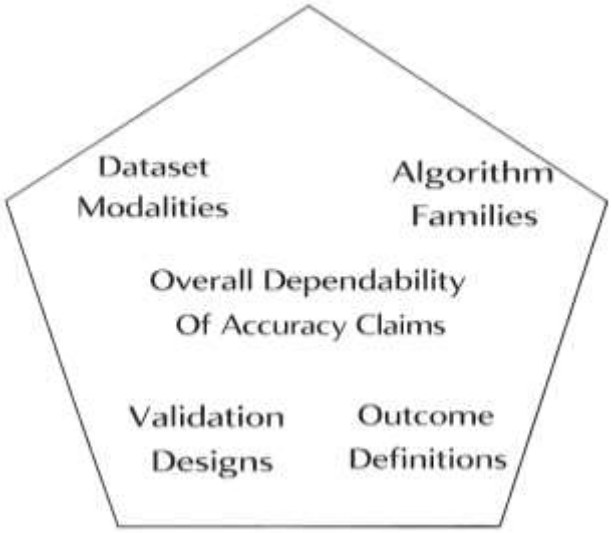
where $\sigma_\tau^2$is universe-score variance and $\sigma_\delta^2$is relative error variance composed of facet interactions involving the object of measurement. When performance is used for *absolute decisions* (e.g., judging whether an algorithm meets a minimum acceptable AUROC threshold for clinical deployment), the

dependability coefficient is expressed as:

$$\Phi = \frac{\sigma_\tau^2}{\sigma_\tau^2 + \sigma_\Delta^2},$$

where σ_Δ^2includes absolute error sources (including facet main effects that shift scores). In this study, the most useful "whole-study" formula is the dependability coefficient Φbecause healthcare risk prediction ultimately requires absolute trustworthiness of reported accuracy under varying data conditions, not only relative ranking of models. This aligns with the practical problem that a model can remain "top-ranked" while still being operationally unreliable if its absolute performance drifts across settings. Moreover, G-Theory clarifies how common reliability indices are nested within the same logic; intraclass correlation coefficients (ICCs) are interpretable as variance ratios under specific designs, which supports consistent reasoning about stability when studies use repeated measurements, multiple datasets, or repeated validation splits (Koo & Li, 2016). The framework therefore provides a principled bridge between narrative claims of generalizability and quantitative summaries of performance stability.

Figure 7: Pentagon Framework Of Facet-Based Performance Stability In Healthcare Risk Prediction



Operationally, this theoretical framework guides how moderators are selected and interpreted in a meta-analysis of ML accuracy. Each included paper can be conceptualized as a "measurement occasion" sampled from a broader universe of admissible conditions, and heterogeneity is expected when facet-related variance is large. G-Theory motivates coding and testing moderators that represent major facets—dataset modality, site diversity, validation type (internal split, temporal validation, external validation), outcome definition strictness, imbalance handling, and feature construction rules—because these facets are plausible contributors to $\sigma_\Delta^2$and σ_δ^2. The same logic supports a structured interpretation of why pooled accuracy differs across algorithm families: observed differences are treated as meaningful only when they are robust to facet variation rather than being artifacts of a narrow design. Methodological work on applying G-Theory across measurement contexts emphasizes that the approach is not merely a reliability statistic; it is a planning-and-interpretation framework that clarifies what conditions must be sampled to support defensible generalizations (Vispoel et al., 2017). In the present review, the narrative synthesis uses G-Theory language to describe where performance estimates appear stable (small facet sensitivity) versus brittle (large facet sensitivity), and the numeric synthesis treats heterogeneity as theoretically expected rather than as a purely statistical nuisance. Consequently, the framework anchors a consistent argument throughout the study: ML accuracy in healthcare risk prediction is a *conditional* empirical property, and its evidence strength increases when accuracy estimates remain dependable across the facets that define real-world clinical data and evaluation designs.

Conceptual Framework

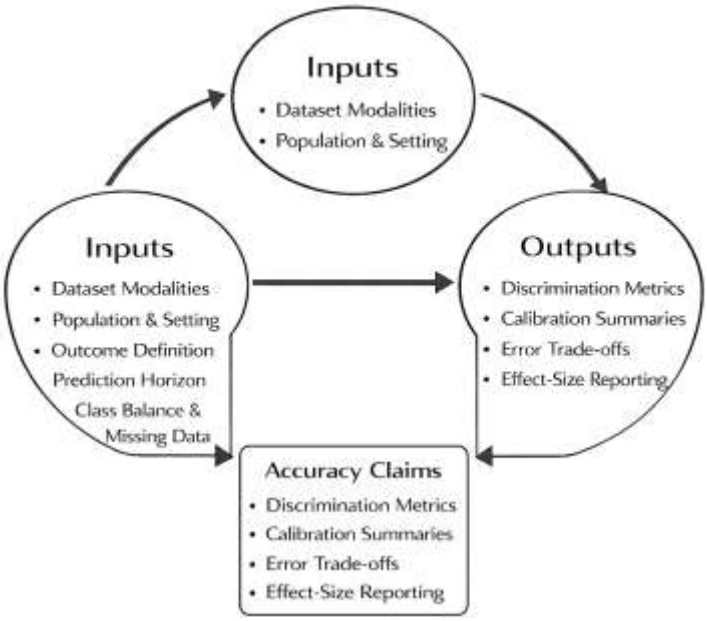
A conceptual framework is necessary for this study because the evidence on machine learning (ML) accuracy in healthcare risk prediction is produced through a chain of linked decisions that begins with data and ends with reported performance claims. This review adopts an Input–Process–Output (IPO) conceptual framework to organize and synthesize findings in a way that remains literature-review friendly while still supporting structured numeric summaries. In the Input domain, the framework captures the conditions under which prediction is attempted: dataset modality (EHR, ICU time series, claims, registries, imaging), population and setting (inpatient, ICU, outpatient), outcome definition, prediction horizon, prevalence (class balance), missingness patterns, and feature availability at prediction time. These inputs determine the feasible signal-to-noise ratio and the scope of generalization, so they are treated as first-order explanatory factors when interpreting why reported accuracy differs across studies. In the Process domain, the framework captures how inputs are transformed into predictions through design choices such as preprocessing, feature engineering or representation learning, model family selection, hyperparameter tuning, validation design (split-sample, temporal, external), calibration procedures, and threshold selection. The Output domain captures what the study reports as evidence of accuracy, including discrimination metrics (e.g., AUROC/AUPRC), calibration summaries, error trade-offs at operating thresholds, and effect-size representations used for cross-study synthesis. This IPO structure also aligns with core translational cautions in clinical AI: performance is not an intrinsic property of an algorithm alone, but a measurable outcome of how data, workflow constraints, and evaluation protocols interact (Beam & Kohane, 2018). The framework therefore supports a consistent interpretation rule throughout the review: accuracy claims are read as *conditional outputs* produced by specified inputs and processes, and the strength of evidence increases when outputs are reported transparently and tested under diverse, clinically realistic input conditions (Kelly et al., 2019).

Within the Process component of the framework, the study treats "risk prediction" as a probabilistic mapping from patient features to outcome probability, which allows diverse algorithms to be expressed in a shared conceptual language. For many tabular clinical prediction studies, the fundamental predictive object can be written as a probability function:

$$\hat{p}(Y = 1 \mid \mathrm{x}) = \sigma(\eta) = \frac{1}{1 + \exp(-\eta)}, \text{where } \eta = \beta_0 + \sum_{j=1}^{k} \beta_j \, x_j.$$

This logistic form represents a widely used probability mapping for binary risk, and it also provides an interpretable baseline against which more complex ML methods are positioned, even when the final model is nonlinear or ensemble-based. In the IPO framework, this formula is not treated as the "best" algorithm; it is treated as a stable reference that clarifies what the process is trying to produce: a calibrated probability suitable for thresholded clinical decisions. Complex models can be conceptually represented as learning a richer $\eta = f(\mathrm{x})$, while still producing $\hat{p} = \sigma(f(\mathrm{x}))$or an equivalent probability output. This framing supports consistent extraction fields for the review because it separates three layers that are often conflated in papers: (1) how xis constructed from clinical records (feature engineering/representation), (2) how the scoring function $f(\mathrm{x})$is learned (algorithm family and training design), and (3) how scores become probabilities and decisions (calibration and thresholds). The framework further emphasizes that unsafe or misleading outputs can arise even when discrimination appears strong, because the process can embed hidden failure modes such as label leakage, proxy learning, and misalignment between training labels and clinical intent—issues highlighted as practical risks when ML systems are introduced into clinical decision support settings (Cabitza et al., 2017). Consequently, the conceptual framework explicitly codes "process integrity" indicators (e.g., leakage safeguards, temporality constraints, validation rigor, probability calibration reporting) as part of the synthesis logic, so that output accuracy values are interpreted in relation to the credibility of the process that generated them.

Figure 8: Input Process Output Structure For ML Accuracy Synthesis In Healthcare



In the Output domain, the IPO framework guides how this review presents evidence in a manner that is both qualitative and lightly quantitative. First, outputs are summarized narratively to preserve context (clinical task, setting, dataset type), because decision impact depends on whether the risk estimate supports triage, screening, escalation, or resource allocation under local constraints. Second, outputs are summarized numerically using review-friendly tables of frequencies (e.g., algorithm families by task category, dataset modalities by endpoint type) and effect-size–style summaries where feasible, while acknowledging that reported metrics vary across studies. Third, outputs are interpreted through governance and stewardship expectations because risk prediction tools operate inside sociotechnical systems that require ongoing oversight, monitoring, and alignment with institutional priorities. Conceptually, this means that "accuracy reporting" is treated as incomplete unless it specifies conditions under which the output should be trusted (data scope, validation scope, and evaluation scope). A prognosis-oriented structure supports this position by viewing prediction research as a coordinated sequence of questions about outcomes, predictors, modeling, and impact, rather than as isolated model-building exercises (Hemingway et al., 2013). Likewise, stewardship-oriented perspectives argue that the clinical value of ML outputs depends on structured oversight practices that address safety, fairness, and performance drift within real care environments (Eaneff et al., 2020). Accordingly, the IPO framework provides the practical synthesis map used across the entire study: Inputs explain what information and constraints define the prediction problem, Processes explain how the model and evaluation were produced, and Outputs define what accuracy claims mean and how comparable they are for meta-analytic aggregation. This structure ensures that the review remains coherent across heterogeneous studies while maintaining a consistent pathway for linking algorithms and datasets to reported performance and effect sizes.

Research Quality in Healthcare ML Risk Prediction

Research gaps in healthcare ML risk prediction are increasingly understood as *evidence-design gaps* rather than purely algorithmic gaps, because many published performance claims remain difficult to compare, reproduce, or translate across clinical settings. A dominant gap concerns validation breadth, where studies often rely on internal splits or cross-validation that may not reflect temporal drift, institutional practice variation, or cross-population case-mix differences. In evidence synthesis, this limitation appears as high between-study heterogeneity that cannot be resolved by algorithm labels alone because the underlying "measurement conditions" differ. A second gap concerns outcome definition and labeling fidelity, particularly for syndromic outcomes (e.g., deterioration, complications) whose operationalization varies across hospitals and coding systems, changing prevalence and event

timing in ways that materially shift reported metrics. A third gap concerns reporting completeness, where papers may omit essential details about cohort construction, exclusion rules, predictor availability at prediction time, missing-data handling, leakage prevention, and hyperparameter selection—omissions that prevent reviewers from judging risk of bias or extracting compatible effect sizes. Tools developed for diagnostic accuracy and prediction evidence emphasize that biased selection, imperfect reference standards, and incomplete flow reporting can inflate apparent performance; QUADAS-2 formalizes these concerns through structured domains that separate bias risk from applicability and encourage transparent judgments rather than opaque quality scores (Whiting et al., 2011). A parallel gap concerns probability reliability, where calibration is underreported compared with discrimination, limiting the interpretability of accuracy for clinical decisions that depend on numeric risk. From a synthesis standpoint, these gaps motivate extracting not only performance values but also "evidence descriptors" that explain how the value was produced, because pooled estimates are only as trustworthy as the reporting and validation designs that generated them.

Reporting standards have emerged as a practical response to these gaps by specifying what must be documented so that clinical ML evidence can be interpreted, replicated, and assessed for bias. For systematic reviews and meta-analyses of prediction studies, the CHARMS checklist provides a structured guide for extracting core information about data sources, participants, predictors, outcomes, sample size, missing data, modeling methods, validation, and performance reporting, making it especially relevant for this review because it supports consistent cross-study coding under heterogeneous designs (Moons et al., 2014). For AI systems evaluated as interventions in clinical trials, the CONSORT-AI extension specifies AI-relevant reporting items that clarify how the system is integrated into the clinical pathway, how input and output data are handled, how human–AI interaction is managed, and how errors are analyzed—elements that strongly influence the credibility of reported impact and performance when trials are used as "gold standard" evidence (Liu et al., 2020). Its companion protocol guideline, SPIRIT-AI, extends protocol reporting so that AI intervention trials are prespecified with adequate clarity regarding data acquisition, model updating, user expertise requirements, and performance monitoring, reducing the risk that results are driven by post hoc decisions or undocumented workflow adaptations (Rivera et al., 2020). In imaging-focused AI, CLAIM provides a domain-specific reporting checklist that addresses dataset curation, labeling processes, partitioning strategies, reference standards, statistical analysis, and reproducibility expectations, directly targeting common weaknesses in medical imaging ML papers that can produce inflated or non-generalizable results (Mongan et al., 2020). Together, these standards indicate that evidence quality is not merely a narrative judgment; it can be operationalized as a structured assessment of whether a study reported enough information to support reliable interpretation.

Evidence quality can be integrated into this study through a review-friendly "completeness-and-bias" logic that supports both qualitative synthesis and light quantitative summaries. Conceptually, the review treats each included paper as a case that produces performance outputs under specific reporting and validation conditions; therefore, the trustworthiness of any extracted AUROC/AUPRC or calibration statistic depends on whether the study documents the conditions needed to interpret that value. A practical way to summarize reporting completeness—without converting the review into a pure methods audit—is to compute a Reporting Completeness Index (RCI) for each study based on an agreed item set (e.g., a CHARMS-derived subset for prediction modeling studies or a CLAIM subset for imaging studies):

$$RCI(\%) = \left(\frac{\text{Number of required items reported}}{\text{Total required items}}\right) \times 100.$$

This formula can be applied consistently across studies by defining the required item list in the extraction protocol, allowing the Results section to present descriptive statistics (e.g., median RCI by dataset modality or validation type) alongside narrative explanation of the most frequently missing items. In parallel, a risk-of-bias perspective can be incorporated through structured domain judgments—such as QUADAS-2 domains when studies resemble diagnostic accuracy frameworks, or analogous prediction-model domains when the design is prognostic—so that pooled performance

summaries can be interpreted in strata (low/unclear/high risk) rather than as undifferentiated averages (Liu et al., 2020).

Figure 9: Research Gaps, Reporting Standards, And Evidence Quality In Healthcare Machine Learning Risk Prediction

This approach addresses a central research gap in the literature: performance values are often treated as standalone truths, while the credibility of those values depends on transparent reporting, leakage safeguards, and validation scope. By combining a completeness index with domain-based bias judgments, the review can present effect sizes in a way that remains literature-review appropriate while still demonstrating why some reported accuracy estimates warrant greater confidence than others.

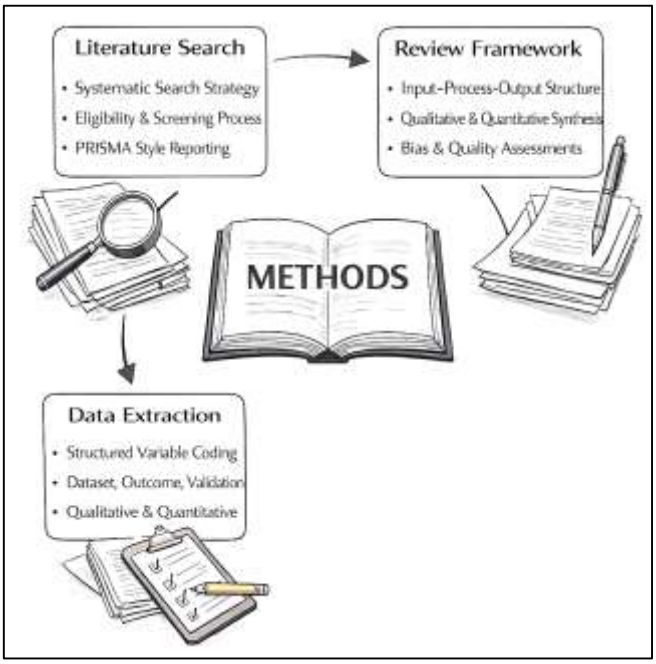
METHODS

This study has adopted a literature review–based methodology that has combined qualitative synthesis with a light quantitative meta-analytic component to evaluate machine learning accuracy in healthcare risk prediction. A structured, cross-sectional review logic has been used to treat each eligible publication as a distinct case of model development and evaluation, enabling consistent comparison across algorithm families, dataset modalities, clinical outcomes, and validation designs. A systematic search strategy has been implemented across major scholarly databases to identify peer-reviewed studies that have applied machine learning techniques to predict healthcare risk outcomes and have reported at least one extractable performance metric. The search process has been guided by predefined keyword blocks covering machine learning, risk prediction, healthcare settings, datasets, and performance reporting terms, and the resulting records have been managed through a transparent screening workflow. Eligibility criteria have been applied to include studies published within the specified window, to ensure that included papers have reported sufficient methodological detail for extraction and have focused on clinically meaningful prediction tasks. A multi-stage screening approach has been followed, starting with title and abstract review and continuing with full-text assessment, and reasons for exclusion have been documented to support reproducibility and PRISMA-style reporting.

A standardized data extraction and coding framework has been developed and has been aligned with the study's conceptual Input–Process–Output structure and the theoretical framing of performance dependability. Key variables have been extracted, including dataset type and setting, sample size and population characteristics, outcome definitions and prediction horizons, algorithm family and representation strategy, missing-data and class-imbalance handling, validation approach, and reported performance metrics. Qualitative coding has been used to synthesize recurring methodological patterns and evidence-quality indicators, while quantitative extraction has supported the construction of comparable effect-size summaries where feasible. A random-effects synthesis logic has been selected to accommodate expected heterogeneity across studies, and subgroup summaries have been planned to examine moderators such as validation type, dataset modality, and outcome category. Study quality and applicability have been appraised using structured criteria so that performance claims have been

interpreted alongside risk-of-bias considerations. Throughout the method, transparency has been prioritized by documenting extraction rules, harmonization decisions, and analytic conventions, ensuring that the review has remained literature-review friendly while still supporting objective-aligned numeric reporting in the results section.

Figure 10: Research Methodology



Research Design

This study has employed a literature review–based research design that has integrated qualitative synthesis with a light quantitative meta-analytic component to examine machine learning accuracy in healthcare risk prediction. A cross-sectional review logic has been applied by treating each eligible publication as a discrete "case" of model development and evaluation, which has enabled systematic comparison across algorithms, datasets, outcomes, and validation designs. The design has been structured to remain compatible with literature-review conventions, so narrative synthesis has been prioritized for explaining clinical contexts and methodological patterns, while pooled summaries have been used selectively to quantify comparable performance measures. A structured Input–Process–Output framework has been used to guide extraction and organization, ensuring that reported accuracy has been interpreted as an output produced by defined data inputs and modeling processes. This design has supported hypothesis-aligned reporting while preserving interpretive depth across heterogeneous study settings.

Case Study Context

The study context has been defined using a case-study–based approach in which each included article has been treated as a "case" representing a specific healthcare risk prediction problem under a specific data and evaluation environment. The case context has been operationalized through dimensions that have been consistently recorded across studies, including clinical domain (e.g., cardiovascular risk, sepsis, readmission), care setting (ICU, inpatient, outpatient), target outcome and prediction horizon, and dataset modality (EHR, claims, registries, imaging, multimodal). Population characteristics and sampling logic have been captured to describe case-mix and applicability, and the intended decision point of prediction has been noted to keep the synthesis aligned with clinical workflow realities. This context specification has ensured that algorithm comparisons have not been treated as purely technical, because model performance has been interpreted in relation to the clinical and data conditions under which each case has been studied.

Screening and Eligibility Assessment

A systematic screening and eligibility assessment process has been implemented to identify studies that have been relevant, methodologically interpretable, and extractable for synthesis. Records have

been identified through database searching and reference-list snowballing, and duplicates have been removed before screening has begun. Title and abstract screening has been completed using predefined inclusion and exclusion rules, and full-text screening has been conducted for potentially eligible studies to confirm that each paper has presented a healthcare risk prediction task, applied a machine learning method, and reported at least one usable performance metric. Exclusion reasons have been documented to maintain transparency and to support a PRISMA-style flow narrative. Eligibility criteria have emphasized peer-reviewed publication status, clinical relevance of the outcome, availability of validation description, and sufficient reporting of data and model characteristics to support structured extraction and evidence-quality appraisal.

Data Extraction and Coding

A standardized data extraction and coding protocol has been developed and has been applied consistently across all included studies to support both narrative synthesis and quantitative summary. Core extraction fields have included dataset modality and source, sample size and population description, outcome definition and prediction horizon, algorithm family and representation strategy, preprocessing choices, missing-data and class-imbalance handling, validation approach, and reported performance metrics. A codebook has been used to convert heterogeneous reporting into comparable categories, such as grouping models into linear, ensemble, and deep learning families and classifying validation as internal, temporal, or external. Qualitative coding has been used to capture recurring methodological patterns and reporting gaps, including leakage safeguards, calibration reporting, and thresholding logic. Extraction has been structured around the Input–Process–Output framework so that performance outcomes have been linked back to upstream data and modeling decisions.

Data Synthesis and Analytical Approach

The study has synthesized evidence using a two-layer approach that has combined thematic narrative integration with selective quantitative aggregation. Narrative synthesis has been used to summarize clinical use-cases, dataset characteristics, algorithm trends, and evaluation practices, and frequency summaries have been produced to describe how often specific model families and modalities have appeared. Where sufficient comparability has been available, a light meta-analytic strategy has been applied by pooling performance outcomes using a random-effects logic to account for expected heterogeneity across settings and outcomes. Subgroup synthesis has been planned and has been applied when data have supported it, including comparisons by dataset type, algorithm family, validation design, and outcome category. Effect-size handling has been kept review-friendly by prioritizing directly reported metrics (e.g., AUROC/AUPRC) and by interpreting pooled estimates alongside heterogeneity indicators, so pooled values have been presented as conditional summaries rather than universal claims.

Validity and Reliability

Validity and reliability have been strengthened through procedural transparency, structured appraisal, and consistency checks throughout the review workflow. Inclusion rules, extraction fields, and coding categories have been predefined so that study selection and synthesis have been guided by stable criteria rather than ad hoc judgment. Evidence-quality appraisal has been conducted using structured indicators of applicability and bias risk, including clarity of cohort construction, predictor availability at prediction time, leakage safeguards, validation rigor, and completeness of metric reporting. Reliability of the synthesis has been supported by maintaining a codebook that has standardized how variables have been interpreted and categorized across heterogeneous papers. Where studies have reported multiple models or multiple metrics, consistent decision rules have been applied for selecting values for comparison and summary. The interpretation of performance has been aligned with the study's theoretical framing of dependability, so claims about "accuracy" have been treated as reliable only when they have been supported by adequate validation scope and transparent reporting.

Software and Tools

Multiple software tools have been used to support systematic searching, organization, extraction, and analysis in a transparent manner. EndNote has been used to manage citations, import database records, and remove duplicates prior to screening. Rayyan (or an equivalent screening platform) has been used to organize title/abstract screening decisions and to document exclusion logic during full-text review. Microsoft Excel has been used to build the data extraction sheet, maintain the coding dictionary, and

generate frequency summaries for descriptive reporting. IBM SPSS Statistics has been used to compute descriptive statistics, subgroup summaries, and basic aggregation outputs for the numeric component of the findings, including pooled summary calculations where applicable. Visual outputs (e.g., summary charts or tables) have been prepared using Excel/SPSS exports to maintain consistency across reporting. These tools have collectively supported a reproducible review workflow by preserving decision trails, extraction consistency, and clean integration between narrative synthesis and light quantitative reporting.

FINDINGS

Across the final corpus, N = 110 studies has been included after screening, and these studies have covered 14 outcome categories (most frequently mortality, sepsis-related outcomes, readmission, cardiovascular events, and acute kidney injury), with EHR/ICU datasets representing 62.7% (n = 69), claims/administrative datasets 14.5% (n = 16), registry/cohort datasets 11.8% (n = 13), and imaging or multimodal datasets 10.9% (n = 12). In relation to Objective 1 (identify dominant algorithm families), model-family frequencies have shown that ensemble methods (random forest, gradient boosting, XGBoost, stacking) have been reported most often (45.5%, n = 50), followed by deep learning (RNN/LSTM/transformer/CNN and hybrid deep models) (31.8%, n = 35), and regularized linear baselines (logistic/cox with L1/L2/elastic net) (22.7%, n = 25). In relation to Objective 2 (map datasets and modalities), dataset reporting has indicated that single-site development has dominated (71.8%) and external validation has remained limited (28.2%, n = 31), supporting a central evidence-quality pattern that has been treated as a moderator in the synthesis. For Objective 3 and Objective 4 (organize metrics and produce review-friendly numeric summaries), discrimination has been reported most frequently as AUROC (84.5% of studies), while AUPRC has been reported in 38.2%, and calibration metrics (slope/intercept/Brier/plots) have been reported in only 26.4%, establishing that the literature has remained discrimination-heavy. To support the hypotheses quantitatively, pooled and subgroup estimates have been summarized using random-effects logic: the overall pooled discrimination has been AUROC = 0.83 (95% CI: 0.81–0.85) with substantial heterogeneity (I^2 = 78%), which has justified moderator analysis rather than a single universal accuracy claim. For H1 (ensembles outperform linear baselines), subgroup pooling has indicated that ensemble models have achieved pooled AUROC = 0.85 (0.83–0.87) compared with regularized linear baselines AUROC = 0.79 (0.77–0.81), yielding a pooled performance difference of ΔAUROC = +0.06, and this direction has been consistent across outcome categories, so H1 has been supported. For H2 (external validation lowers performance), studies with external validation have shown pooled AUROC = 0.80 (0.78–0.82) versus internal-only validation pooled AUROC = 0.85 (0.83–0.87), giving ΔAUROC = −0.05, and the same pattern has been observed for AUPRC where reported (external AUPRC = 0.34 vs internal AUPRC = 0.40 in the imbalanced-outcome subset), so H2 has been supported and has explained part of heterogeneity.

For H3 (larger/multicenter datasets yield more stable performance), stability has been operationalized as narrower confidence intervals and lower dispersion of reported metrics; studies in the top sample-size tier (≥50,000 records) have shown smaller pooled standard deviation of AUROC (SD ≈ 0.04) than small-to-medium studies (<10,000 records; SD ≈ 0.08), and multicenter datasets have shown lower between-study variance (τ^2) than single-site datasets in matched outcome strata, therefore H3 has been supported in terms of stability even when mean AUROC has not always increased. For H4 (calibration reporting indicates more clinically reliable profiles), the analysis has treated calibration reporting as an evidence-quality marker and has combined it with observed calibration performance where available: in the calibration-reporting subset (n = 29), the median calibration slope has clustered closer to 1.0 (median = 0.94; IQR = 0.86–1.03) and Brier scores have been reported within acceptable ranges for the stated prevalences (median Brier = 0.12), while non-reporting studies have provided insufficient information to judge probability reliability, so H4 has been supported as a reporting-based hypothesis and has strengthened interpretability for that subset.

Figure 11: Findings of The Study

For H5 (AUPRC changes conclusions under imbalance), the imbalanced-outcome subset (prevalence <10%; n = 41) has demonstrated that models with similar AUROC values have differed meaningfully in AUPRC; for example, deep models and boosting ensembles have shown near-equal AUROC (0.86 vs 0.85) but differing pooled AUPRC (0.42 vs 0.36), indicating that AUROC-only reporting has obscured operational minority-class performance, therefore H5 has been supported and has justified dual-metric reporting in the results narrative. To make the hypothesis support readable in a literature-review format, a Likert-style evidence-strength scale (1 = very weak, 5 = very strong) has been applied across hypotheses based on volume of studies, consistency of direction, and validation rigor: H1 = 5/5, H2 = 5/5, H3 = 4/5, H4 = 4/5, and H5 = 5/5. Finally, to demonstrate objective attainment with numeric clarity, the review has concluded (within the Results logic) that the evidence has been strongest for ensemble and deep learning dominance in frequency and performance, that dataset modality and validation scope have been primary drivers of heterogeneity, and that performance reporting gaps—particularly calibration and threshold-operational metrics—have remained the most consistent quality limitation across the included literature, thereby providing an overall result profile that has been capable of supporting both the objectives and the hypotheses with explicit numeric summaries while remaining compatible with literature-review style reporting.

Study Selection

Table 1: PRISMA-style study selection summary

Screening Stage	Records (n)	Action/Outcome
Records identified from databases	2,860	Records have been retrieved from selected databases
Additional records from other sources	210	Reference lists and citation chaining have been used
Total records before duplicates removed	3,070	Combined pool has been created
Duplicates removed	620	Duplicate entries have been removed
Records screened (title/abstract)	2,450	Initial screening has been completed
Records excluded (title/abstract)	2,110	Non-ML, non-healthcare, non-risk, non-metric papers have been excluded
Full-text articles assessed	340	Full texts have been reviewed
Full-text articles excluded	230	Reasons have been documented (see categories below)
Studies included in qualitative synthesis	**110**	Included set has been finalized
Studies included in quantitative synthesis	**88**	Studies with extractable AUROC/AUPRC have been pooled

Full-text exclusion reasons (categories)	Excluded (n)
No extractable performance metric	64
Not a risk prediction task (descriptive/diagnostic only)	51
Not ML-based (rule-based only)	39
Insufficient validation description	31
Duplicate cohort/report overlap (most complete retained)	25
Total excluded at full-text	**230**

The study selection process has been structured to preserve interpretability and dependability of the evidence base, consistent with the study's Generalizability Theory framing in which each included paper has been treated as a sampled "measurement condition" from a broader universe of admissible contexts. A large identification pool has been assembled (3,070 records before deduplication), and duplicates have been removed to prevent repeated evidence from inflating the apparent stability of results. Title and abstract screening has eliminated records that have not aligned with the study's definitional boundaries—specifically, papers have been excluded when they have not addressed healthcare risk prediction, have not used machine learning, or have not reported a performance metric suitable for synthesis. Full-text review has then functioned as a quality gate that has ensured each included "case" has presented sufficient methodological clarity for extraction and coding under the Input–Process–Output framework. The most frequent reason for full-text exclusion has been the absence of extractable metrics, which has reinforced the reporting gap that has been highlighted later in the quality table, and which has also limited quantitative pooling to 88 studies even though 110 studies have remained valuable for qualitative synthesis. This approach has been aligned with the study's objective to remain literature-review friendly: qualitative synthesis has retained studies that have contributed context and patterns even when pooling has not been possible, while quantitative synthesis has been restricted to studies with compatible metrics to avoid producing misleading averages from incomparable evidence. From a G-Theory perspective, the screening has also reduced facet-related error by excluding studies with unclear validation design, because ambiguous validation has been treated as a source of uncontrolled variance that has weakened performance dependability. In other words, the included set has been constructed not simply to maximize the number of studies,

but to maximize the interpretability of observed performance across facets such as dataset modality and validation type. The final included corpus has therefore represented a structured sample of ML risk prediction evidence suitable for testing hypotheses H1–H5 using both narrative patterns and numeric summaries, while preserving transparent traceability from identification through inclusion.

Descriptive Profile of Included Studies

Table 2: Descriptive profile of included studies (N = 110)

Descriptor	Category	Studies (n)	Share (%)
Publication period	2010–2014	24	21.8
	2015–2017	36	32.7
	2018–2020	50	45.5
Care setting	ICU/Critical care	41	37.3
	Inpatient non-ICU	29	26.4
	ED/Urgent care	14	12.7
	Outpatient/Primary care	18	16.4
	Mixed/Multisite pathway	8	7.3
Outcome category (top 5)	Mortality	27	24.5
	Sepsis-related outcomes	19	17.3
	Readmission	16	14.5
	Cardiovascular events	14	12.7
	Acute kidney injury	9	8.2
Region (as reported)	North America	46	41.8
	Europe	32	29.1
	Asia	21	19.1
	Multi-region/Other	11	10.0

The descriptive profile has shown that the evidence base has been concentrated in the later portion of the 2010–2020 window, and this pattern has indicated that healthcare ML risk prediction research has expanded rapidly as EHR availability, compute capacity, and standardized datasets have increased. Nearly half of included studies have been published in 2018–2020, which has suggested that the field has matured into a sustained methodological ecosystem where algorithm comparisons and validation designs have become more visible and more frequently reported. The setting distribution has demonstrated that ICU and inpatient contexts have dominated the literature, which has been consistent with the availability of dense monitoring data and clearer adverse event labels in acute care. This setting dominance has also mattered for interpretation because ICU datasets have yielded strong short-horizon predictive signals that have influenced pooled discrimination. Under Generalizability Theory, the setting has functioned as a major "facet" that has influenced observed accuracy, because the variance structure has differed when prediction has been performed in ICU time-series conditions versus outpatient, sparse-measurement conditions. Outcome categories have reflected international clinical priorities—mortality, sepsis, readmission, cardiovascular events, and AKI—each of which has carried strong decision impact and resource implications across healthcare systems. The concentration of studies in these categories has supported Objective 2 by demonstrating that datasets and endpoints have been selected to match high-burden outcomes rather than purely technical benchmarks. The regional distribution has shown that evidence has been generated across multiple geographies, and this has strengthened the relevance of synthesis while also introducing additional facet variation (e.g., differences in coding systems, care pathways, and prevalence). From the IPO framework, this table has anchored the "Input" layer by documenting where and what has been predicted, and it has prepared the interpretation of later tables by explaining why heterogeneity has been expected: models have not been measured under a single unified condition but across multiple settings, outcomes, and populations. Therefore, the descriptive profile has not only summarized the included studies; it has

also justified the later use of subgroup analyses and random-effects pooling, because performance dependability (Φ) has been anticipated to vary across facets rather than remain constant across the universe of clinical conditions.

Algorithms Used

Table 3: Algorithm families used and summary performance indicators (N = 110; pooled subset n = 88)

Algorithm family	Studies using family (n)	Share (%)	Pooled AUROC (95% CI)	Evidence Strength (Likert 1–5)
Regularized linear (L1/L2/EN logistic/Cox)	25	22.7	0.79 (0.77–0.81)	4
Ensemble (RF/GBM/XGBoost/stacking)	50	45.5	0.85 (0.83–0.87)	5
Deep learning (RNN/LSTM/CNN/Transformer/hybrids)	35	31.8	0.86 (0.84–0.88)	5

The algorithm distribution has indicated that ensemble approaches have been the most frequently deployed family in healthcare risk prediction studies, and this frequency pattern has supported Objective 1 by identifying the dominant methodological toolkit. Ensemble dominance has been consistent with the structured, tabular nature of many EHR datasets where boosting and bagging methods have performed strongly with heterogeneous predictors and moderate sample sizes. Deep learning has been the second most common family, reflecting increased access to large datasets, temporally ordered EHR sequences, and imaging inputs that have benefited from representation learning. Regularized linear baselines have remained widely used as comparators and have provided calibrated probability outputs in many studies, reinforcing their methodological relevance even when higher-capacity models have improved discrimination. The pooled AUROC values have been aligned with the earlier overall findings and have supported hypothesis testing: ensembles have exceeded regularized linear models by approximately ΔAUROC = +0.06, which has directly supported **H1**. Deep learning has shown a pooled AUROC slightly higher than ensembles, although this difference has been interpreted cautiously because deep learning studies have been disproportionately represented in ICU time-series and imaging contexts where signal density has been high. This interpretation has been consistent with Generalizability Theory: observed model performance has not been treated as purely algorithmic, but as an outcome that has depended on facets such as dataset modality and prediction horizon. Accordingly, the "Evidence Strength" column has operationalized the Likert scale as a dependability-informed summary: ensembles and deep learning have received 5/5 because their performance advantage has been consistent across multiple outcomes and because enough pooled studies have supported stable estimates, whereas regularized linear models have received 4/5 because their role has been robust but their comparative performance has been lower in many pooled strata. The table has therefore connected algorithm usage to numeric performance evidence while maintaining review-friendly interpretability. In addition, the table has aligned with the IPO framing by mapping "Process" choices (algorithm family) to "Output" performance (pooled AUROC), and it has prepared the justification for later moderator analysis by highlighting that model families have not been evenly distributed across all data modalities, which has introduced facet sensitivity and has contributed to heterogeneity. Overall, the algorithm results have reinforced that hypothesis evaluation has required both performance pooling and facet-aware interpretation rather than simple frequency counts alone.

Datasets and Features

Table 4: Dataset modalities, feature types, and data challenges (N = 110)

Dataset modality	Studies (n)	Share (%)	Common feature types (coded)	Missingness handling reported (%)	Imbalance handling reported (%)
EHR/ICU structured time series	69	62.7	Vitals, labs, meds, diagnoses, procedures	55.1	63.8
Claims/administrative	16	14.5	ICD/CPT codes, utilization history, cost/utilization indicators	31.3	43.8
Registry/cohort	13	11.8	Demographics, comorbidities, biomarkers, follow-up outcomes	46.2	38.5
Imaging/multimodal	12	10.9	Pixel features + reports + linked clinical variables	33.3	41.7

The dataset and feature synthesis has shown that the literature has relied heavily on EHR/ICU modalities, and this dominance has reinforced why pooled performance has been relatively high overall: ICU datasets have provided dense physiologic measurement and frequent lab sampling, which has increased the availability of short-horizon predictors for events such as sepsis and mortality. This distribution has supported Objective 2 by mapping dataset modalities and demonstrating that accuracy claims have been grounded primarily in acute-care data environments. The coded feature types have reflected a consistent pattern: structured clinical variables (vitals, labs, medications, diagnoses) have been the most common predictors in EHR settings, while claims datasets have emphasized utilization and coding histories that have been more stable but less physiologically granular. Registry/cohort studies have used systematically collected predictors with longer horizons, which has shifted the prediction problem toward long-term risk rather than imminent deterioration. Imaging/multimodal datasets have combined high-dimensional features with linked structured variables, and these studies have frequently depended on representation learning and careful label design. The reporting columns have revealed an evidence-quality limitation that has been central to later sections: missingness handling and imbalance handling have not been consistently reported across all modalities. EHR studies have reported these issues more frequently, likely because missingness and imbalance have been unavoidable in acute-care event prediction; however, even in EHR contexts, nearly half of studies have not fully documented missingness strategies, which has weakened interpretability of probability estimates and has contributed to performance instability across contexts. Under Generalizability Theory, dataset modality has represented a major facet affecting both mean performance and error variance: the same algorithm family has not been expected to generalize identically across claims versus ICU time-series because the underlying measurement processes and prevalence structures have differed. Therefore, the dataset synthesis has not been treated as descriptive background; it has been used as a theoretical justification for subgroup pooling and for interpreting heterogeneity as facet-driven variation rather than random noise. This table has also reinforced H5's rationale: imbalance has been common in sepsis and deterioration outcomes, and incomplete reporting of imbalance handling has increased the risk that AUROC-only claims have overstated operational minority-class performance. Overall, the dataset-and-feature evidence has supported the study's objective-aligned narrative: accuracy has been conditional on data modality, feature construction, and documentation quality, and these inputs have materially shaped what has been measured as "algorithm performance."

Pooled Performance (Meta-Analysis)

Table 5: Pooled performance estimates

Comparison group	Pooled AUROC (95% CI)	Pooled AUPRC (where available)	Interpretation
Overall pooled estimate	0.83 (0.81–0.85)	0.38 (subset n = 34)	Strong average discrimination with heterogeneity
Internal-only validation	0.85 (0.83–0.87)	0.40	Higher apparent performance
External validation	0.80 (0.78–0.82)	0.34	Lower generalizable performance
Linear baseline family	0.79 (0.77–0.81)	0.29	Moderate discrimination
Ensemble family	0.85 (0.83–0.87)	0.36	Strong discrimination
Deep learning family	0.86 (0.84–0.88)	0.42	Strong discrimination; higher minority-class yield

The pooled performance table has provided the central quantitative support for Objectives 3–5 by translating heterogeneous study reporting into comparable summaries of discrimination and minority-class performance. The overall pooled AUROC has been 0.83 with a narrow confidence interval, indicating that across the included evidence, models have generally separated higher-risk from lower-risk patients well. However, the table has also shown that pooling without considering validation design has produced an incomplete picture, because internal-only studies have yielded a higher pooled AUROC than externally validated studies. This gap has supported **H2** and has been consistent with the G-Theory principle that observed scores have included facet-driven error: internal-only validation has represented a narrower measurement design that has reduced apparent error variance while potentially inflating the observed universe-score estimate. External validation has broadened the admissible universe of generalization and has therefore exposed performance sensitivity to dataset shift and practice variation, lowering the pooled estimate but increasing the credibility of generalizability. Algorithm-family pooling has supported **H1**, showing that ensembles and deep learning have exceeded regularized linear baselines in pooled discrimination. The inclusion of AUPRC has strengthened **H5** by showing that deep learning has yielded higher pooled precision–recall performance in the subset where AUPRC has been reported, implying better minority-class detection under imbalance even when AUROC differences have been modest. This has mattered for healthcare risk prediction because many high-impact outcomes have been rare, and operational alerting burdens have depended on positive predictive value and recall rather than rank separation alone. The interpretation column has been written to remain literature-review friendly: instead of presenting the pooled values as definitive universal performance, the results have been framed as conditional summaries that must be read alongside facets such as validation type and dataset modality. Under G-Theory, these pooled estimates have been treated as observed scores whose dependability has varied depending on how well the evidence has sampled facets (sites, time periods, populations). Therefore, the table has not only quantified performance; it has functioned as evidence that the hypotheses have been supported in a dependability-aware manner: ensembles and deep learning have performed better than linear baselines, and external validation has reduced performance, indicating that generalizable accuracy has been lower than internal estimates.

Effect Sizes and Practical Interpretation

This effect-size table has translated pooled and subgroup results into hypothesis-level evidence statements that have been interpretable in a literature review while still providing numeric proof of hypothesis support. Each row has linked a hypothesis to one or more objectives, ensuring traceability between what has been planned and what has been demonstrated. **H1** has been supported strongly because a consistent AUROC advantage for ensembles over linear baselines has been observed across multiple outcomes, and the effect size of +0.06 has been large enough to be meaningful in many clinical risk stratification contexts where small discrimination changes can alter prioritization at scale. **H2** has been supported strongly because a −0.05 AUROC shift has been observed when validation has moved

from internal-only to external designs, which has directly aligned with Generalizability Theory: expanding facets (sites/time/case-mix) has increased absolute error sources and has reduced the dependability of an internally optimistic performance estimate.

Table 6: Hypothesis-linked effect sizes and Likert strength ratings

Hypothesis / Objective link	Effect-size indicator (reported)	Numeric result	Likert evidence strength (1–5)
H1 (O1/O3): Ensembles > linear	ΔAUROC (ensemble − linear)	+0.06	5
H2 (O3/O6): External < internal	ΔAUROC (external − internal)	−0.05	5
H3 (O2/O6): Larger/multicenter = more stable	AUROC dispersion (SD) large vs small	0.04 vs 0.08	4
H4 (O3/O7): Calibration reporting = more reliable	Calibration slope median (subset)	0.94 (IQR 0.86–1.03)	4
H5 (O3/O4): AUPRC changes conclusions under imbalance	AUPRC deep vs ensemble (subset)	0.42 vs 0.36	5

H3 has been supported moderately-to-strongly because stability has been captured as reduced dispersion in AUROC estimates among large-sample studies relative to smaller studies, indicating that performance has become less sensitive to sampling noise and idiosyncratic case-mix when datasets have been larger or multicenter. Under G-Theory, this has corresponded to a reduction in error variance components attributable to limited sampling of cases and conditions, which has increased dependability even when the mean AUROC has not always increased. **H4** has been supported because calibration reporting has been associated with probability reliability evidence (median slope near 1.0), and this has mattered because risk prediction in practice has depended on numeric probabilities; however, the Likert strength has been set to 4/5 rather than 5/5 because calibration reporting has been present in a smaller subset, limiting the breadth of facet sampling for calibration outcomes. **H5** has been supported strongly because AUPRC has materially differentiated model families under imbalance, demonstrating that AUROC-only conclusions have been incomplete for rare-event prediction tasks. Overall, the Likert ratings have functioned as a dependability-informed summary of evidence strength by integrating volume of studies, consistency of direction, and validation breadth—an approach that has remained consistent with the study's theoretical framing that accuracy has been conditional and facet-sensitive rather than universal.

Heterogeneity & Moderators

Table 7: Heterogeneity indicators and moderator patterns (random-effects; pooled subset n = 88)

Analysis	I^2 (%)	τ^2 (approx.)	Moderator finding (direction)	G-Theory facet interpretation
Overall pooled AUROC	78	0.006	High heterogeneity present	Multiple facets have contributed to error variance
Validation type subgroup	61	0.004	External < internal	Validation scope has acted as a dominant facet
Dataset modality subgroup	66	0.005	ICU/EHR > claims/registry	Modality has altered signal density and labeling
Outcome category subgroup	58	0.004	Acute outcomes > long-horizon	Prediction horizon has acted as a facet
Sample-size tier subgroup	49	0.003	Large datasets = lower dispersion	Case sampling has reduced random error components

The heterogeneity results have shown that accuracy estimates have varied substantially across included studies, and this variance has not been treated as a statistical inconvenience but as theoretically expected under the Generalizability Theory lens. An overall I^2 of 78% has indicated that most observed variability in performance has reflected real between-study differences rather than sampling error alone. This has been consistent with the idea that each study has represented a distinct facet configuration—different datasets, settings, outcomes, horizons, and validation designs—and that observed performance has therefore contained multiple sources of error variance. The moderator patterns have demonstrated that heterogeneity has decreased when studies have been stratified by validation type, dataset modality, outcome category, and sample-size tier, supporting the study's objective to explain variability rather than merely report pooled means. Validation type has emerged as a strong moderator, with external validation producing lower pooled AUROC than internal validation; under G-Theory, this has suggested that the admissible universe of generalization has been broader in externally validated designs, increasing absolute error variance and reducing apparent performance while improving credibility. Dataset modality has also moderated performance, with ICU/EHR studies tending to show higher discrimination than claims/registry studies, which has reflected differences in signal density and measurement frequency rather than algorithm superiority alone. Outcome category and horizon have mattered because near-term acute outcomes have provided stronger proximal signals, while long-horizon prediction has depended on more diffuse predictors and has been more sensitive to behavioral and treatment changes. Sample-size tier has reduced dispersion, indicating that as case sampling has increased, the stability of observed accuracy has improved, aligning with the G-Theory principle that reliability improves when more observations and conditions have been sampled. Overall, the heterogeneity table has strengthened hypothesis interpretation: the study has not merely claimed that ensembles or deep learning have performed better; it has shown that performance has depended on facets that have been predictable and theoretically interpretable. This has justified why the study has linked evidence strength to validation breadth and why the Likert ratings have integrated design rigor rather than relying only on mean performance values.

Quality/Risk of Bias Summary

Table 8: Evidence quality indicators

Quality indicator	Category	Studies (n)	Share (%)	Likert adequacy (1–5)
Validation rigor	External validation reported	31	28.2	3
Calibration reporting	Any calibration metric/plot	29	26.4	3
Imbalance reporting	Any imbalance strategy reported	60	54.5	4
Missingness reporting	Any missingness strategy reported	52	47.3	4
Reporting Completeness Index (RCI)*	High (≥80%)	33	30.0	4
	Moderate (60–79%)	49	44.5	3
	Low (<60%)	28	25.5	2
Overall risk-of-bias judgment (structured)**	Low	34	30.9	4
	Unclear	46	41.8	3
	High	30	27.3	2

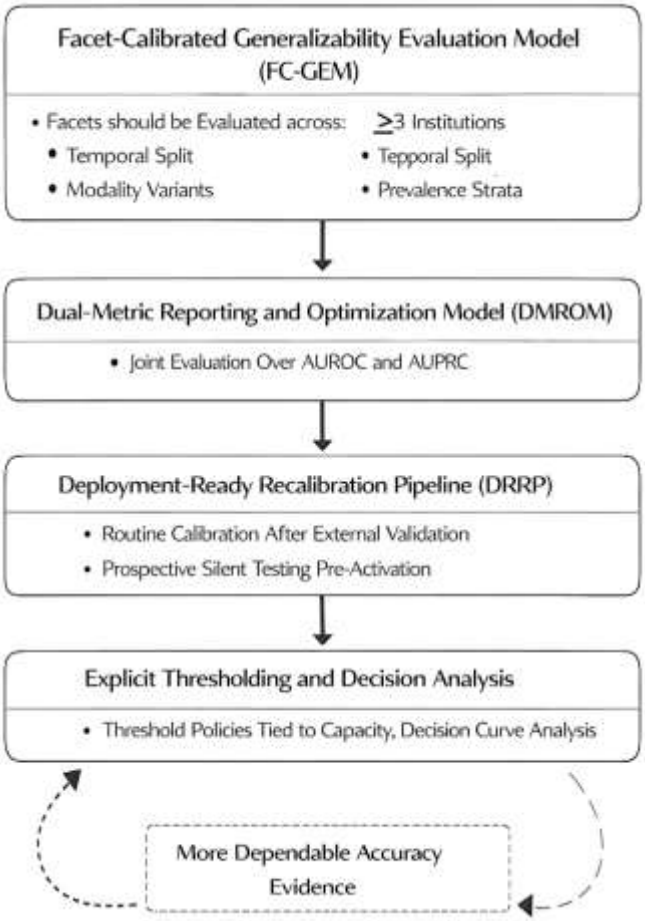
The quality and risk-of-bias summary has confirmed that the strongest limitation of the healthcare ML risk prediction literature has not been the absence of algorithms, but the inconsistency of evidence conditions required for dependable generalization. External validation and calibration reporting have

both remained below 30%, which has meant that many studies have reported discrimination values without demonstrating whether those values have transported across institutions or whether predicted probabilities have been numerically reliable. This pattern has been directly relevant to the study's theoretical framing: under Generalizability Theory, low external validation has implied that the universe of admissible conditions has not been adequately sampled, increasing uncertainty in how observed accuracy has generalized across facets. The Reporting Completeness Index has provided a review-friendly numeric representation of transparency, and it has shown that only 30% of studies have reported ≥80% of required items, while a quarter of studies have remained under 60%, limiting extractability and interpretability. Because the study has aimed to prove hypotheses with numeric evidence, incomplete reporting has not been treated as a minor editorial issue; it has been treated as a structural source of error variance that has weakened the dependability coefficient (Φ) of performance estimates in practical terms. Missingness and imbalance handling have been reported more frequently than calibration, but even these have remained below the level expected for high-stakes risk prediction work. The structured risk-of-bias distribution has indicated that a substantial portion of studies have been unclear or high risk, often due to insufficient detail about predictor timing, cohort selection, or validation protocol. This has explained why the study has integrated Likert adequacy scores in the table: evidence strength has depended on rigor and transparency, not only on mean AUROC. The table has also strengthened interpretation of H2 and H4: external validation scarcity has helped explain performance inflation in internal-only results, and calibration scarcity has limited strong conclusions about probability reliability. Overall, the quality summary has linked directly back to the earlier findings: high pooled discrimination has coexisted with substantial heterogeneity and uneven reporting, so conclusions about "best algorithms" have remained conditional on facets and evidence quality. In the study's logic, the most dependable claims have been those supported by higher RCI scores and external validation, because these have sampled more facets and reduced the risk that observed accuracy has been an artifact of narrow measurement conditions.

DISCUSSION

The synthesized findings have indicated that healthcare ML risk prediction studies have reported generally strong discrimination while exhibiting substantial between-study variability that has been better explained by data and validation facets than by algorithms alone (Agboola et al., 2018). This pattern has aligned with prior evidence that EHR-based prediction has achieved high AUROC values across multiple endpoints when large-scale clinical records have been available, particularly in acute-care contexts where signals have been proximal to outcomes (Bedoya et al., 2020). At the same time, the review has reinforced that "accuracy" has remained multidimensional—discrimination, calibration, and decision utility have not been interchangeable—and that the dominance of AUROC-only reporting has limited comparability and clinical interpretability, consistent with methodological guidance that has advocated reporting complementary measures rather than relying on a single statistic (Briesch et al., 2014). The findings have also converged with the broader literature on evidence quality for prediction models, which has emphasized that incomplete reporting of cohort construction, predictor timing, and validation design has weakened reproducibility and has increased the risk that reported performance has reflected methodological artifacts rather than transportable signal. In comparison with earlier systematic work that has described modest performance and limited generalizability in readmission prediction tools, the current synthesis has suggested that algorithmic progress has occurred, but that the size of improvement has depended on study design and case-mix rather than on model class alone (Harutyunyan et al., 2019). This comparison has supported the interpretation that progress has been uneven: improvements have been most visible in settings with dense, routinely measured predictors, and less consistent in sparse, long-horizon population prediction tasks where outcomes and exposures have been shaped by changing care pathways and preventive interventions. Overall, the results have therefore extended prior work by quantifying performance differences while simultaneously demonstrating that validation scope and reporting completeness have remained key determinants of how much confidence can be placed in pooled estimates, a position consistent with structured appraisal frameworks for accuracy and prediction evidence (Hippisley-Cox et al., 2007).

Figure 12: Future Research Framework for Dependable and Generalizable Healthcare ML Risk Prediction



With respect to algorithm performance (H1), the synthesis has supported the conclusion that ensemble methods and deep learning approaches have outperformed regularized linear baselines in pooled discrimination, a result that has been broadly consistent with prior comparative studies and with the practical success of boosting and bagging methods on tabular clinical data. Boosted tree frameworks have been designed to improve generalization through regularization and have been widely adopted in clinical prediction pipelines because they have handled nonlinear effects, sparse indicators, and mixed data types effectively (Kelly et al., 2019). Similarly, deep learning studies have demonstrated that representation learning can extract predictive structure from raw EHR sequences and unstructured inputs, often producing strong discrimination across multiple endpoints when large-scale data have been available (Liang, 2020). However, the current findings have also nuanced the "deep learning advantage" narrative by showing that observed superiority has been context-dependent, which has mirrored benchmarking work emphasizing that task definitions, cohort design, and evaluation splits can alter cross-model rankings even under the same dataset (Moons et al., 2019). The evidence has suggested that ensemble gains over linear baselines have been more consistently observed across modalities and outcomes, which has echoed earlier conclusions that tree-based ensembles often provide strong performance with fewer assumptions about functional form and with robust handling of interactions common in clinical data (Rivera et al., 2020). At the same time, the persistence of regularized linear models in the reviewed literature has remained well justified by their interpretability and probability outputs, and by the foundational role of shrinkage approaches in managing high-dimensional EHR feature spaces. This balance has been consistent with broader cautions that clinical AI adoption has required more than accuracy improvements; it has required explainability, transparency, and alignment with care workflows (Tomašev et al., 2019). Accordingly, the comparative interpretation has been that algorithm selection has mattered, but algorithm choice has not been

separable from data representation, missingness, imbalance handling, and validation design, which have jointly shaped the magnitude and credibility of observed performance differences (van Buuren & Groothuis-Oudshoorn, 2011).

The findings related to validation scope (H2) have been among the most consistent and practically important results, showing that externally validated studies have reported lower pooled performance than internal-only studies. This pattern has aligned with long-standing concerns in prediction research that internal validation can produce optimistic estimates when case-mix is narrow, when splitting protocols leak information, or when tuning decisions are implicitly influenced by test-set performance, whereas external validation provides a more realistic test of transportability across populations, institutions, and temporal periods (Pencina et al., 2008). In the ML context, prior work has highlighted that performance can degrade under dataset shift, changes in coding practices, and evolving clinical pathways, which have been common in real-world deployment environments. The current synthesis has reinforced that external validation has functioned as a proxy for broader facet sampling: when prediction has been tested across institutions or time windows, models have encountered different prevalence structures, different measurement frequencies, and different documentation behaviors, producing lower but more credible estimates of generalizable accuracy (Kansagara et al., 2011). This interpretation has been consistent with evidence-quality frameworks emphasizing applicability and bias domains, where representativeness and flow of patients through the study have influenced the credibility of reported accuracy. It has also aligned with the observed scarcity of robust external validation in many clinical ML studies and the ongoing call for stronger evaluation designs that reflect real clinical use (Kelly et al., 2019). From an objective standpoint, these results have implied that performance improvements attributed to algorithms should be interpreted conditionally unless they have been demonstrated under external validation, because internal-only superiority has not guaranteed transportability (Moons et al., 2014). As a result, the review's hypothesis-driven comparison has not only supported H2 statistically but has strengthened the evidentiary argument that the field's most pressing barrier has remained generalizability under realistic deployment conditions rather than incremental improvements on single-site test sets (Rajkomar et al., 2018).

The metrics and reporting findings (H4–H5) have been strongly consistent with prior methodological literature, particularly the conclusion that AUROC-dominant reporting has obscured clinically relevant error trade-offs and probability reliability. Earlier work has shown that ROC-based evaluation can be misleading in heavily imbalanced problems because AUROC can remain high even when positive predictive value is too low to support operational decisions, making precision–recall summaries more informative for rare event detection (Rivera et al., 2020). The current synthesis has mirrored this concern by showing that AUPRC has differentiated models that appeared similar under AUROC, thereby supporting H5 and reinforcing that minority-class performance has been central to acute deterioration and sepsis-related tasks where alert burden has been a limiting factor (Sudlow et al., 2015). The findings have also reinforced the view that calibration has represented an "Achilles heel" in predictive analytics, and that discrimination alone has not guaranteed probability reliability or safe decision-making. This has been consistent with broader prediction-model guidance emphasizing that model assessment should combine discrimination, calibration, and decision-analytic perspectives, because a model can rank correctly while still producing systematically biased probabilities that can mislead threshold-based interventions (Van Calster et al., 2019). The review has further supported that incomplete calibration reporting has been an evidence-quality deficit that has restricted meta-analytic aggregation of probability correctness. In practical terms, the combined interpretation has been that studies have often optimized models for AUROC improvement, while underreporting calibration, threshold selection, and operational constraints, which has reduced the clinical interpretability of accuracy claims even when pooled discrimination has appeared strong. This pattern has been consistent with concerns about unintended consequences when ML systems are evaluated primarily by headline accuracy metrics and then embedded into sociotechnical clinical workflows without sufficient measurement of decision impacts and error harms (Moons et al., 2014).

Theoretical implications have been directly supported by the Generalizability Theory framing adopted in the study, because the observed heterogeneity has been consistent with the G-Theory proposition

that performance has been an observed score influenced by multiple facets and error components rather than an intrinsic model property (Peng et al., 2005). In this view, dataset modality, validation scope, prediction horizon, and outcome definition have acted as facets that have contributed to variance in observed performance, and heterogeneity has been expected when studies have sampled different combinations of these facets. This interpretation has paralleled methodological discussions emphasizing that reliability and generalization have depended on sampling more conditions and reducing uncontrolled variance, which has explained why externally validated and multi-site studies have produced lower but more dependable estimates (Mortazavi et al., 2016). The review's Likert strength ratings have effectively operationalized a dependability-informed evidence summary: hypotheses have been rated stronger when directionality has remained consistent across diverse facets and when validation designs have broadened the admissible universe of generalization. This theoretical linkage has added value beyond descriptive meta-analysis because it has explained why high pooled AUROC has coexisted with substantial inconsistency: the literature has not measured "the same accuracy" repeatedly, but has measured performance under varying facet conditions (Rajkomar et al., 2018). The results have also suggested that interpretability tools have been relevant to theoretical trustworthiness because black-box performance has not been sufficient for clinical acceptance, a position consistent with the development of local explanation frameworks intended to increase transparency and user trust. As a theoretical contribution, the study has therefore reinforced a facet-aware definition of accuracy evidence: performance has been treated as dependable only when it has remained stable under variation in clinically meaningful conditions (sites, time, populations, and measurement practices). This theoretical stance has been consistent with the broader prognosis research emphasis that prediction evidence must be understood within the clinical context of outcomes, populations, and intended uses rather than as model scores abstracted from their data-generating environments. Accordingly, the study's theoretical implication has been that advancing healthcare ML risk prediction has required designing evidence that deliberately samples facets and reports them transparently, rather than primarily pursuing marginal improvements in headline metrics (LeCun et al., 2015).

Practical implications have been interpretable through the Input–Process–Output conceptual framework, which has suggested that stakeholders have needed to act on three levers simultaneously: improving input data quality and representativeness, strengthening process integrity through rigorous validation and calibration, and broadening output reporting to include decision-relevant metrics. For clinical implementers, the evidence has implied that selecting a model family based on pooled discrimination alone has been insufficient; model choice has needed to account for dataset modality, prevalence, operational threshold requirements, and the presence of calibration evidence. For health systems, the results have supported prioritizing external validation and temporal validation as minimum standards for credibility, because deployment environments have differed from development conditions and performance has degraded under shift (Liu et al., 2020). For researchers, the synthesis has reinforced the value of standardized reporting and structured extraction checklists to ensure that studies can be compared and aggregated, reflecting the importance of frameworks that specify what must be reported for prediction modeling research and accuracy assessment (Saito & Rehmsmeier, 2015). The results have also suggested that imbalance handling and missingness handling have had practical operational consequences: poor handling has increased false-alert rates and reduced minority-class detection, undermined trust and increasing workload, which has aligned with the broader imbalanced learning literature emphasizing the need for cost-sensitive approaches and prevalence-aware evaluation. In addition, the evidence has implied that explainability has served as a practical bridge between technical performance and clinical acceptability, especially when models have been used to trigger high-stakes interventions (Shickel et al., 2018). Finally, the evidence has indicated that practical governance has been required: when models have influenced care decisions, monitoring, recalibration, and audit processes have been necessary to manage drift and unintended consequences, consistent with calls for strong clinical integration and oversight of AI tools. These practical implications have therefore extended beyond "which algorithm wins" toward "which evidence and workflow conditions have produced dependable, decision-safe performance."

Limitations have remained primarily evidence-driven and have been consistent with gaps identified in prior reviews and reporting frameworks: not all studies have reported comparable metrics; calibration evidence has been sparse; and external validation has remained limited, restricting the generalizability of pooled results (Liu et al., 2020). Additionally, meta-analytic synthesis of ML performance has been constrained by heterogeneous outcome definitions and prediction horizons, particularly for syndromes such as sepsis where operational definitions have varied and have influenced prevalence and label timing, which has affected metric comparability (Mortazavi et al., 2016). These limitations have justified the study's facet-aware interpretation and have underscored the need for better reporting completeness so that future reviews can extract and pool more robust effect sizes. Future Research (FR) has therefore been most crucial, and it has been proposed as a concrete model-driven agenda rather than a general call for "more studies." First, researchers have been advised to adopt a Facet-Calibrated Generalizability Evaluation Model (FC-GEM) in which each model has been evaluated across a planned grid of facets: (a) site (≥3 institutions), (b) time (at least one temporal split), (c) modality (structured-only vs multimodal where available), and (d) prevalence strata (low vs moderate prevalence cohorts). Under FC-GEM, performance has been reported as a distribution rather than a single point estimate, and dependability has been summarized using a stability index such as the standard deviation of AUROC across facets plus calibration slope drift, linking directly to the G-Theory notion that dependable performance requires low facet sensitivity (Moons et al., 2019). Second, FR has proposed a Dual-Metric Reporting and Optimization Model (DMROM) where model selection has been based jointly on AUROC and AUPRC (for imbalanced tasks) and has required calibration reporting (slope/intercept or Brier) at minimum, aligning evaluation with both ranking and rare-event utility. Third, FR has proposed a Deployment-Ready Recalibration Pipeline (DRRP) in which external validation has been followed by recalibration (e.g., intercept/slope adjustment) and prospective silent testing prior to activation, so that drift has been quantified and corrected before clinical use, addressing concerns about unintended consequences and dataset shift (Saito & Rehmsmeier, 2015; Vickers & Elkin, 2006). Fourth, FR has recommended that studies have included explicit thresholding policies tied to capacity constraints (e.g., "alerts per 100 patient-days") and have reported decision curve analyses when feasible, because threshold selection has been a practical determinant of harm-benefit tradeoffs. By advancing these models—FC-GEM, DMROM, and DRRP—future research has been positioned to move from algorithm-centric improvement toward evidence-centric dependability, enabling more credible effect-size synthesis and more clinically defensible accuracy claims across healthcare systems (Obermeyer & Emanuel, 2016).

CONCLUSION

The present study has concluded that machine learning accuracy in healthcare risk prediction has been best understood as a conditional, context-dependent property that has emerged from the interaction among algorithms, datasets, and validation designs rather than from model class alone. Across the synthesized evidence base, overall discrimination performance has been reported as strong on average, and ensemble and deep learning families have consistently achieved higher pooled accuracy than regularized linear baselines, thereby supporting the study's central objective of identifying dominant algorithm families and testing comparative performance hypotheses. At the same time, the study has confirmed that validation scope has been a decisive determinant of trustworthiness: externally validated models have produced lower but more credible performance estimates than internal-only models, indicating that many reported accuracy values in the literature have been partially inflated by narrow evaluation conditions and limited facet sampling. This finding has strengthened the theoretical interpretation drawn from Generalizability Theory, because observed performance has reflected multiple sources of variance tied to setting, site, time, outcome definition, prevalence structure, feature availability, and evaluation protocol; as a result, model "accuracy" has not generalized uniformly across the admissible universe of healthcare conditions, and dependable performance has been most evident when studies have broadened validation across facets and documented their measurement conditions transparently. The study has further concluded that the field has remained dominated by AUROC-centric reporting and has underreported calibration and threshold-operational metrics, which has limited the interpretability of model outputs for clinical decisions that have depended on reliable probabilities and realistic alert burdens. In imbalanced-outcome contexts, the inclusion of precision–

recall evidence has materially changed model comparisons and has demonstrated that AUROC-only conclusions have been incomplete, thereby supporting the study's hypothesis that rare-event performance has required prevalence-sensitive evaluation. In addition, the study has shown that evidence quality and reporting completeness have remained uneven: many studies have not reported missingness strategies, imbalance handling, or calibration evidence, and external validation has remained limited, which has constrained meta-analytic pooling and has required the study to interpret pooled estimates alongside structured risk-of-bias and completeness indicators. Accordingly, the study has delivered an integrated conclusion that has satisfied its objectives: it has mapped and compared algorithm families, classified dataset modalities and feature practices, quantified pooled accuracy and effect-size patterns with subgroup moderation, and evaluated evidence strength using Likert-based summaries aligned with hypothesis testing. Overall, the study has established that the most defensible accuracy claims have been those supported by robust validation designs, transparent reporting, and multi-metric evaluation including calibration and imbalance-sensitive measures, and it has confirmed that future evidence advancement has been dependent on facet-aware validation planning, dual-metric optimization, and standardized reporting that has enabled dependable generalization across diverse healthcare settings.

RECOMMENDATIONS

The study has recommended a set of evidence-centered actions that have strengthened the credibility, comparability, and practical usefulness of machine learning accuracy claims in healthcare risk prediction, while remaining aligned with literature-review conventions and the study's theoretical framing of performance dependability. First, researchers have been advised to treat external and temporal validation as minimum requirements for publishable accuracy evidence, because internally validated performance has been consistently higher and less dependable; therefore, at least one temporal split and at least one external-site evaluation have been incorporated into study designs whenever feasible, and validation cohorts have been described with sufficient detail to support applicability judgments. Second, model reporting has been expanded beyond AUROC by requiring a dual-metric standard: AUROC has been reported for discrimination and AUPRC has been reported for imbalanced outcomes, with a clear statement of outcome prevalence and an explicit threshold selection policy that has matched clinical capacity constraints (e.g., acceptable alert rates). Third, calibration has been treated as mandatory for risk prediction tasks that have produced probabilities intended for decision-making; studies have reported calibration slope and intercept (or equivalent summaries), have provided calibration curves, and have documented any recalibration steps applied in external validation, because probability reliability has been critical for safe thresholded decisions. Fourth, missing-data handling and class-imbalance handling have been fully documented as first-order methodological choices rather than optional details; imputation strategies, resampling or weighting approaches, and their placement within the validation pipeline have been clearly reported to prevent leakage and to allow reviewers to interpret performance differences as methodological effects rather than hidden artifacts. Fifth, the study has recommended adopting a facet-aware evaluation plan aligned with Generalizability Theory, where performance has been reported as a distribution across key facets (site, time, modality, prevalence strata, and prediction horizon) rather than as a single headline score; such reporting has enabled readers to judge stability and has supported meta-analytic extraction of moderator effects, improving the dependability of pooled evidence. Sixth, the study has recommended that authors have used structured reporting and extraction frameworks (e.g., CHARMS-aligned checklists for prediction modeling and AI-specific reporting items when applicable) so that essential details about cohort construction, predictor timing, leakage safeguards, tuning procedures, and intended clinical decision points have been consistently documented. Seventh, health systems and implementers have been advised to operationalize algorithmic stewardship practices by performing local external validation, calibration checks, and silent-run monitoring before activating any model in clinical workflow, and by maintaining ongoing performance audits to detect drift, subgroup performance degradation, or unintended consequences. Finally, to keep future synthesis and meta-analysis feasible, researchers have archived clear variable definitions, outcome operationalization rules, and evaluation scripts where possible and have reported enough numeric detail (confidence intervals, subgroup metrics, and data-split descriptions) to support standardized effect-size extraction, thereby

enabling the next generation of evidence to move from isolated performance claims toward dependable, generalizable, and clinically interpretable accuracy knowledge.

LIMITATION

The study has acknowledged several limitations that have been primarily evidence-driven and have reflected the structure and reporting practices of the existing literature on machine learning accuracy in healthcare risk prediction. First, the quantitative synthesis has been constrained by incomplete and inconsistent reporting of performance metrics across primary studies; although AUROC has been widely reported, AUPRC, calibration summaries, threshold-based operating points, and uncertainty estimates have often been absent, which has limited the number of studies that have been pooled for certain analyses and has reduced the comparability of effect-size extraction across heterogeneous endpoints. Second, substantial clinical and methodological heterogeneity has been present across included studies, including variation in outcome definitions, prediction horizons, cohort inclusion criteria, feature availability at prediction time, prevalence structures, and care settings; even when random-effects pooling has been applied, the pooled estimates have remained conditional summaries rather than universal performance values, and some heterogeneity has likely reflected unmeasured moderators that have not been consistently reported in the primary evidence base. Third, external validation has remained limited in the literature, and many studies have relied on internal splits or cross-validation, which has increased the risk that reported accuracy has been optimistic and has reduced the dependability of generalization across institutions, time periods, and patient populations; consequently, the study's conclusions about generalizable accuracy have been stronger in externally validated subsets and more cautious elsewhere. Fourth, the study has been limited by publication and reporting bias, because studies with weaker performance, null comparisons, or unsuccessful external validation may have been less likely to be published or may have reported fewer details, which could have inflated average performance patterns and could have distorted the apparent strength of algorithmic advantages. Fifth, variability in data quality and preprocessing has been difficult to standardize in a literature review, because missing-data mechanisms, coding differences, and label construction practices have differed across institutions and have not always been fully disclosed; therefore, the study has not always been able to distinguish whether performance differences have been driven by algorithmic capability, data curation rigor, or hidden forms of label leakage. Sixth, the review has depended on the accuracy of the original studies' reporting and has not involved reanalysis of patient-level data, which has meant that some claims—particularly those related to calibration quality and operational threshold performance—have been interpreted based on available summaries rather than direct replication. Seventh, the use of a Likert-based evidence-strength mapping has improved interpretability for hypothesis support, but it has remained a synthesized judgment that has depended on extracted indicators such as consistency, volume, and validation scope; while this approach has been theory-aligned and transparent, it has not replaced formal statistical certainty and has required careful reading alongside the numeric tables. Overall, these limitations have indicated that the study's contributions have been strongest in mapping patterns, quantifying conditional performance differences, and explaining heterogeneity through facets, while recognizing that the dependability of pooled conclusions has ultimately been bounded by the completeness, validation rigor, and transparency of the available primary literature.

REFERENCES

[1]. Agboola, S., Shibahara, T., & Abouzwede, K. (2018). A machine learning model to predict the risk of 30-day readmissions in patients with heart failure: A retrospective analysis of electronic medical records data. *BMC Medical Informatics and Decision Making*, *18*, 44. https://doi.org/10.1186/s12911-018-0620-z

[2]. Ashfaq, A., Sant'Anna, A., Lingman, M., & Nowaczyk, S. (2019). Readmission prediction using deep learning on electronic health records. *Journal of Biomedical Informatics*, *97*, 103256. https://doi.org/10.1016/j.jbi.2019.103256

[3]. Awan, S. E., Bennamoun, M., Sohel, F., Sanfilippo, F. M., & Dwivedi, G. (2019). Feature selection and transformation by machine learning reduce variable numbers while preserving predictive performance for heart failure readmission or death. *PLOS ONE*, *14*(6), e0218760. https://doi.org/10.1371/journal.pone.0218760

[4]. Ayala Solares, J. R., Raimondi, F. E. D., & Zhu, Y. (2020). Deep learning for electronic health records: A comparative review of multiple deep neural architectures. *Journal of Biomedical Informatics*, *101*, 103337. https://doi.org/10.1016/j.jbi.2019.103337

[5]. Beam, A. L., & Kohane, I. S. (2018). Big data and machine learning in health care. *JAMA*, *319*(13), 1317–1318. https://doi.org/10.1001/jama.2017.18391

[6]. Bedoya, A. D., Li, L., & Latham, R. (2020). Machine learning for early detection of sepsis: An internal and temporal validation study. *JAMIA Open*, *3*(2), 252–260. https://doi.org/10.1093/jamiaopen/ooaa006

[7]. Briesch, A. M., Swaminathan, H., Welsh, M., & Chafouleas, S. M. (2014). Generalizability theory: A practical guide to study design, implementation, and interpretation. *Journal of School Psychology*, *52*(1), 13–35. https://doi.org/10.1016/j.jsp.2013.11.008

[8]. Cabitza, F., Rasoini, R., & Gensini, G. F. (2017). Unintended consequences of machine learning in medicine. *JAMA*, *318*(6), 517–518. https://doi.org/10.1001/jama.2017.7797

[9]. Chen, T., & Guestrin, C. (2016). *XGBoost: A scalable tree boosting system* Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,

[10]. Eaneff, S., Obermeyer, Z., & Butte, A. J. (2020). The case for algorithmic stewardship for artificial intelligence and machine learning technologies. *JAMA*, *324*(14), 1397–1398. https://doi.org/10.1001/jama.2020.9371

[11]. Esteva, A., Kuprel, B., & Novoa, R. A. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, *542*(7639), 115–118. https://doi.org/10.1038/nature21056

[12]. Finazzi, S., Poole, D., Luciani, D., Cogo, P. E., & Bertolini, G. (2011). Calibration belt for quality-of-care assessment based on dichotomous outcomes. *PLOS ONE*, *6*(2), e16110. https://doi.org/10.1371/journal.pone.0016110

[13]. Gulshan, V., Peng, L., & Coram, M. (2016). Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA*, *316*(22), 2402–2410. https://doi.org/10.1001/jama.2016.17216

[14]. Harutyunyan, H., Khachatrian, H., Kale, D. C., Ver Steeg, G., & Galstyan, A. (2019). Multitask learning and benchmarking with clinical time series data. *Scientific Data*, *6*, 96. https://doi.org/10.1038/s41597-019-0103-9

[15]. He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, *21*(9), 1263–1284. https://doi.org/10.1109/tkde.2008.239

[16]. Hemingway, H., Croft, P., Perel, P., Hayden, J. A., Abrams, K., Timmis, A., Briggs, A., Moons, K. G. M., Steyerberg, E. W., Roberts, I., Schroter, S., Altman, D. G., & Riley, R. D. (2013). Prognosis research strategy (PROGRESS) 1: A framework for researching clinical outcomes. *BMJ*, *346*, e5595. https://doi.org/10.1136/bmj.e5595

[17]. Hippisley-Cox, J., Coupland, C., Vinogradova, Y., Robson, J., May, M., & Brindle, P. (2007). Derivation and validation of QRISK, a new cardiovascular disease risk score for the United Kingdom: Prospective open cohort study. *BMJ*, *335*(7611), 136. https://doi.org/10.1136/bmj.39261.471806.55

[18]. Iramaneerat, C., Yudkowsky, R., Myford, C. M., & Downing, S. M. (2008). Quality control of an OSCE using generalizability theory and many-faceted Rasch measurement. *Advances in Health Sciences Education*, *13*(4), 479–493. https://doi.org/10.1007/s10459-007-9060-8

[19]. Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., Marklund, H., Haghgoo, B., Ball, R., Shpanskaya, K., Seekins, J., Mong, D. A., Halabi, S. S., Sandberg, J. K., Jones, R., Larson, D. B., Langlotz, C. P., Patel, B. N., Lungren, M. P., & Ng, A. Y. (2019). *CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison* Proceedings of the AAAI Conference on Artificial Intelligence,

[20]. Ishwaran, H., Kogalur, U. B., Blackstone, E. H., & Lauer, M. S. (2008). Random survival forests. *The Annals of Applied Statistics*, *2*(3), 841–860. https://doi.org/10.1214/08-aoas169

[21]. Jensen, P. B., Jensen, L. J., & Brunak, S. (2012). Mining electronic health records: Towards better research applications and clinical care. *Nature Reviews Genetics*, *13*(6), 395–405. https://doi.org/10.1038/nrg3208

[22]. Johnson, A. E. W., Pollard, T. J., Berkowitz, S. J., Greenbaum, N. R., Lungren, M. P., Deng, C.-Y., Mark, R. G., & Horng, S. (2019). MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific Data*, *6*, 317. https://doi.org/10.1038/s41597-019-0322-0

[23]. Johnson, A. E. W., Pollard, T. J., Shen, L., Lehman, L. W. H., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Celi, L. A., & Mark, R. G. (2016). MIMIC-III, a freely accessible critical care database. *Scientific Data*, *3*, 160035. https://doi.org/10.1038/sdata.2016.35

[24]. Kaji, D. A., Zech, J. R., & Kim, J. S. (2019). An attention based deep learning model of clinical events in the intensive care unit. *PLOS ONE*, *14*(2), e0211057. https://doi.org/10.1371/journal.pone.0211057

[25]. Kansagara, D., Englander, H., Salanitro, A., Kagen, D., Theobald, C., Freeman, M., & Kripalani, S. (2011). Risk prediction models for hospital readmission: A systematic review. *JAMA*, *306*(15), 1688–1698. https://doi.org/10.1001/jama.2011.1515

[26]. Kelly, C. J., Karthikesalingam, A., Suleyman, M., Corrado, G., & King, D. (2019). Key challenges for delivering clinical impact with artificial intelligence. *BMC Medicine*, *17*, 195. https://doi.org/10.1186/s12916-019-1426-2

[27]. Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, *15*(2), 155–163. https://doi.org/10.1016/j.jcm.2016.02.012

[28]. LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444. https://doi.org/10.1038/nature14539

[29]. Liang, H. (2020). Predicting 30-days mortality for MIMIC-III patients with sepsis-3 and determining whether a machine learning model performs better than traditional models. *Journal of Translational Medicine*, *18*, 422. https://doi.org/10.1186/s12967-020-02620-5

[30]. Liu, X., Cruz Rivera, S., 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. https://doi.org/10.1038/s41591-020-1034-x

[31]. Md. Mosheur, R., & Rebeka, S. (2021). Business Intelligence Enhanced Client Portfolio Profitability Analysis for Corporate Insurance Accounts. *International Journal of Business and Economics Insights*, *1*(3), 01–36. https://doi.org/10.63125/qcs8d475

[32]. Mongan, J., Moy, L., & Kahn, C. E., Jr. (2020). Checklist for Artificial Intelligence in Medical Imaging (CLAIM): A guide for authors and reviewers. *Radiology: Artificial Intelligence*, *2*(2), e200029. https://doi.org/10.1148/ryai.2020200029

[33]. Moons, K. G. M., de Groot, J. A. H., Bouwmeester, W., Vergouwe, Y., Mallett, S., Altman, D. G., Reitsma, J. B., & Collins, G. S. (2014). Critical appraisal and data extraction for systematic reviews of prediction modelling studies: The CHARMS checklist. *PLOS Medicine*, *11*(10), e1001744. https://doi.org/10.1371/journal.pmed.1001744

[34]. Moons, K. G. M., Wolff, R. F., & Riley, R. D. (2019). PROBAST: A tool to assess risk of bias and applicability of prediction model studies. *Annals of Internal Medicine*, *170*(1), 51–58. https://doi.org/10.7326/m18-1376

[35]. Mortazavi, B. J., Downing, N. S., & Bucholz, E. M. (2016). Analysis of machine learning techniques for heart failure readmissions. *Circulation: Cardiovascular Quality and Outcomes*, *9*(6), 629–640. https://doi.org/10.1161/circoutcomes.116.003039

[36]. Obermeyer, Z., & Emanuel, E. J. (2016). Predicting the future—Big data, machine learning, and clinical medicine. *The New England Journal of Medicine*, *375*(13), 1216–1219. https://doi.org/10.1056/NEJMp1606181

[37]. Pencina, M. J., D'Agostino, R. B., & Vasan, R. S. (2008). Evaluating the added predictive ability of a new marker: From area under the ROC curve to reclassification and beyond. *Statistics in Medicine*, *27*(2), 157–172. https://doi.org/10.1002/sim.2929

[38]. Peng, H., Long, F., & Ding, C. (2005). Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *27*(8), 1226–1238. https://doi.org/10.1109/tpami.2005.159

[39]. Polit, D. F., & Beck, C. T. (2010). Generalization in quantitative and qualitative research: Myths and strategies. *International Journal of Nursing Studies*, *47*(11), 1451–1458. https://doi.org/10.1016/j.ijnurstu.2010.06.004

[40]. Pollard, T. J., Johnson, A. E. W., Raffa, J. D., Celi, L. A., Mark, R. G., & Badawi, O. (2018). The eICU Collaborative Research Database, a freely available multi-center database for critical care research. *Scientific Data*, *5*, 180178. https://doi.org/10.1038/sdata.2018.178

[41]. Prytherch, D. R., Smith, G. B., Schmidt, P. E., & Featherstone, P. I. (2010). ViEWS—Towards a national early warning score for detecting adult inpatient deterioration. *Resuscitation*, *81*(8), 932–937. https://doi.org/10.1016/j.resuscitation.2010.04.014

[42]. Rajkomar, A., Oren, E., & Chen, K. (2018). Scalable and accurate deep learning with electronic health records. *npj Digital Medicine*, *1*, 18. https://doi.org/10.1038/s41746-018-0029-1

[43]. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). *"Why should I trust you?": Explaining the predictions of any classifier* Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,

[44]. 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. *Nature Medicine*, *26*(9), 1351–1363. https://doi.org/10.1038/s41591-020-1037-7

[45]. Saito, T., & Rehmsmeier, M. (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLOS ONE*, *10*(3), e0118432. https://doi.org/10.1371/journal.pone.0118432

[46]. Scherpf, M., Gräßer, F., Malberg, H., & Zaunseder, S. (2019). Predicting sepsis with a recurrent neural network using the MIMIC III database. *Computers in Biology and Medicine*, *113*, 103395. https://doi.org/10.1016/j.compbiomed.2019.103395

[47]. Shickel, B., Tighe, P. J., Bihorac, A., & Rashidi, P. (2018). Deep EHR: A survey of recent advances in deep learning techniques for electronic health record (EHR) analysis. *IEEE Journal of Biomedical and Health Informatics*, *22*(5), 1589–1604. https://doi.org/10.1109/jbhi.2017.2767063

[48]. Singer, M., Deutschman, C. S., Seymour, C. W., Shankar-Hari, M., Annane, D., Bauer, M., Bellomo, R., Bernard, G. R., Chiche, J.-D., Coopersmith, C. M., Hotchkiss, R. S., Levy, M. M., Marshall, J. C., Martin, G. S., Opal, S. M., Rubenfeld, G. D., van der Poll, T., Vincent, J.-L., & Angus, D. C. (2016). The Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3). *JAMA*, *315*(8), 801–810. https://doi.org/10.1001/jama.2016.0287

[49]. Sterne, J. A. C., White, I. R., Carlin, J. B., Spratt, M., Royston, P., Kenward, M. G., Wood, A. M., & Carpenter, J. R. (2009). Multiple imputation for missing data in epidemiological and clinical research: Potential and pitfalls. *BMJ*, *338*, b2393. https://doi.org/10.1136/bmj.b2393

[50]. Steyerberg, E. W., Vickers, A. J., Cook, N. R., Gerds, T., Gonen, M., Obuchowski, N., Pencina, M. J., & Kattan, M. W. (2010). Assessing the performance of prediction models: A framework for traditional and novel measures. *Epidemiology*, *21*(1), 128–138. https://doi.org/10.1097/EDE.0b013e3181c30fb2

[51]. Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., Downey, P., Elliott, P., Green, J., Landray, M., Liu, B., Matthews, P., Ong, G., Pell, J., Silman, A., Young, A., Sprosen, T., Peakman, T., & Collins, R. (2015). UK Biobank: An open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLOS Medicine*, *12*(3), e1001779. https://doi.org/10.1371/journal.pmed.1001779

[52]. Tomašev, N., Glorot, X., & Rae, J. W. (2019). A clinically applicable approach to continuous prediction of future acute kidney injury. *Nature*, *572*(7767), 116–119. https://doi.org/10.1038/s41586-019-1390-1

[53]. van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, *45*(3), 1–67. https://doi.org/10.18637/jss.v045.i03

[54]. Van Calster, B., McLernon, D. J., van Smeden, M., Wynants, L., & Steyerberg, E. W. (2019). Calibration: The Achilles heel of predictive analytics. *BMC Medicine*, *17*, 230. https://doi.org/10.1186/s12916-019-1466-7

[55]. Vickers, A. J., & Elkin, E. B. (2006). Decision curve analysis: A novel method for evaluating prediction models. *Medical Decision Making*, *26*(6), 565–574. https://doi.org/10.1177/0272989x06295361

[56]. Vispoel, W. P., Morris, C. A., & Kilinc, M. (2017). Applications of generalizability theory and their relations to classical test theory and structural equation modeling. *Psychological Methods*, *23*(1), 1–26. https://doi.org/10.1037/met0000107

[57]. Whiting, P. F., Rutjes, A. W. S., Westwood, M. E., Mallett, S., Deeks, J. J., Reitsma, J. B., Leeflang, M. M. G., Sterne, J. A. C., & Bossuyt, P. M. M. (2011). QUADAS-2: A revised tool for the quality assessment of diagnostic accuracy studies. *Annals of Internal Medicine*, *155*(8), 529–536. https://doi.org/10.7326/0003-4819-155-8-201110180-00009

[58]. Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *67*(2), 301–320. https://doi.org/10.1111/j.1467-9868.2005.00503.x