



Integrating Large-Scale Data Analytics for Cardiovascular Disease Prediction: A Scoping Review

Salam Bani Hani¹, Muayyad M. Ahmad²

¹Department of Nursing, Faculty of Nursing, Irbid National University, Irbid, Jordan

²Department of Clinical Nursing, School of Nursing, The University of Jordan, Amman, Jordan

Objectives: This scoping review synthesizes literature on the integration of large-scale data analytics for cardiovascular disease (CVD) prediction, aiming to provide insights that support the adoption of predictive analytics for improved prevention and early detection in healthcare. **Methods:** Searches were conducted in Medline (PubMed), EBSCO, Google Scholar, and Wiley Online Library. Medical Subject Headings (MeSH) search terms included: large-scale data, big data, cardiovascular diseases, prediction, machine-learning algorithms, artificial intelligence, and mortality. The search covered the period from 2020 to 2024. **Results:** Of 262 retrieved articles, 16 were included. Three main themes were identified: large-scale data analysis techniques and machine-learning algorithms; applications of machine-learning algorithms and artificial intelligence in predicting cardiovascular diseases; and the role of integrating large-scale data in disease prediction to improve the quality of care. **Conclusions:** While machine learning provides considerable opportunities for predicting CVD outcomes, limitations remain. Machine-learning approaches are not always the most appropriate option, particularly in basic research where causal relationships between variables may be more critical than optimized predictions. To ensure fair and effective healthcare outcomes, issues related to bias, data quality, ethical concerns, and practical implementation must be addressed. Overcoming these challenges will require interdisciplinary collaboration, methodological refinement, and further research.

Keywords: Cardiovascular Diseases, Data Analytics, Machine Learning, Predictive Value of Tests, Big Data

Submitted: March 4, 2025

Revised: May 16, 2025

Accepted: July 3, 2025

Corresponding Author

Salam Bani Hani

Department of Nursing, Faculty of Nursing, Irbid National University, Irbid, Jordan. Tel: +962-785577701, E-mail: s.banihani@inu.edu.jo (<https://orcid.org/0000-0003-0848-5615>)

This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (<http://creativecommons.org/licenses/by-nc/4.0/>) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

© 2025 The Korean Society of Medical Informatics

I. Introduction

Globally, cardiovascular disorders account for the highest number of deaths [1]. As part of the epidemiological transition, they have become the fourth-leading cause of death, increasing in prevalence across many countries, including Western Europe, North America, Australia, and New Zealand, while displaying substantial regional variation in mortality rates [2]. Consequently, cardiovascular diseases have overtaken infections, also referred to as communicable diseases, as the dominant health burden [3]. Ischemic heart disease (IHD) is now the leading cause of death worldwide and in Western developed and industrialized nations [4].

Large-scale data analytics enable the integration, analysis,

and interpretation of massive datasets from diverse sources, thereby transforming both clinical practice and cardiovascular disease (CVD) research [5]. For instance, natural language processing can extract unstructured electronic health record (EHR) data, such as clinical notes, to enhance risk stratification models, while geospatial analytics can link CVD prevalence to environmental exposures. The Framingham Heart Study has been extensively used in predictive analytics to refine risk scores such as the atherosclerotic cardiovascular disease (ASCVD) score by incorporating demographic and lifestyle factors [6,7].

A review study by Traina et al. [8] focusing on the Middle East and North Africa reported that CVD is the leading cause of death in Arab countries, accounting for 14.3% of cases. Younger adults in this region showed significantly higher rates of myocardial infarction (MI) presentation compared with global averages [9]. The study also highlighted elevated prevalence rates of established risk factors, including diabetes, hypertension, hyperlipidemia, smoking, obesity, and sedentary lifestyles.

The high prevalence of cardiovascular diseases underscores the need for comprehensive and standardized guidelines to support accurate diagnosis and effective treatment [10,11]. Several diagnostic methods are available, ranging from non-invasive procedures such as electrocardiography and exercise stress testing to invasive techniques including blood tests and cardiac catheterization [12].

Large-scale data analysis in CVD prediction involves the processing of complex datasets, such as genetic profiles, medical imaging, EHR, and wearable device monitoring. Artificial intelligence approaches—including machine learning (ML) and deep learning—are employed to facilitate early diagnosis, stratify patient risk, and identify novel risk factors.

The terms “large-scale data” and “big data” are closely related, both referring to the collection, processing, and analysis of vast datasets from heterogeneous sources [13,14]. Their overlap lies in the shared objective of generating valuable insights to inform decision-making, with big data often serving as the foundation for large-scale data analysis in domains such as healthcare, finance, and technology [15]. In this review, the two terms are used interchangeably. Importantly, the integration of large-scale data analytics in disease prediction represents a transformative advance in healthcare, offering the potential to improve patient outcomes [13,16].

Several studies have demonstrated that combining multiple categories of input data, rather than relying on single-domain data, enhances the predictive performance of CVD models [17,18]. The most accurate models typically integrate

data from most or all categories, particularly within ML or deep learning frameworks. These models often achieve areas under the curve (AUCs) exceeding 0.90 in retrospective validations, though they also pose challenges related to data integration, missingness, and interpretability [4,19].

By leveraging big data, healthcare providers can generate more accurate predictions regarding disease onset, progression, and patient-specific risk factors. This facilitates earlier interventions and supports more personalized care strategies [20,21]. This scoping review synthesizes existing evidence on large-scale data analytics in CVD prediction, identifies research gaps, guides future investigations, informs clinical practice, and supports policy development, thereby advancing precision medicine and improving patient care.

II. Methods

1. Protocol and Registration

This scoping review followed the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) Extension for Scoping Reviews guidelines. The protocol for this study was registered with the Open Science Framework on September 9, 2024 (<https://osf.io/9xty2/>).

2. Search Strategy

A comprehensive literature search was conducted for studies published between 2020 and 2025. The following databases were searched: Medline (PubMed), EBSCO, Google Scholar, and Wiley Online Library. Medical Subject Headings (MeSH) search terms included: large-scale data, big data, ischemic heart diseases, prediction, ML algorithms, artificial intelligence (AI), and mortality. Boolean operators “AND” and “OR” were applied to combine terms, which allowed for refining the search. The initial search was performed on August 1, 2024, and the final search on October 6, 2024.

3. Inclusion and Exclusion Criteria

The inclusion criteria were as follows: (1) English-language publications; (2) focus on the use of AI for CVD prediction to improve healthcare outcomes; (3) mixed-method, qualitative, or quantitative research designs; and (4) publication between 2020 and 2024. Exclusion criteria were studies unrelated to AI in CVD prediction or healthcare outcomes, as well as editorials, letters, reports, non-English publications, and research published prior to 2020.

4. Study Selection

A total of 242 articles were retrieved from electronic data-

bases, and 20 additional records were identified from other sources. After 50 duplicates were removed, 75 articles were excluded following title and abstract screening for relevance, language, publication year, and connection to CVD. During full-text review, 37 articles were excluded because they were non-research publications (e.g., editorials, letters, or reports) or did not address large-scale data use. Ultimately, 16 articles were included in the final review (Figure 1).

5. Data Collection

The Johns Hopkins Nursing Evidence-Based Practice Research Appraisal Tool was used to evaluate each publication for relevance [22]. Team members discussed disagreements regarding study eligibility in scheduled meetings. Patterns and conclusions emerging from the included studies were deliberated by the research team.

6. Risk of Bias in Individual Studies

The methodological quality of selected papers was independently assessed using the Quality Checklist for the Assessment Tool (QAT) developed by Thomas et al. [23] and adapted from the Effective Practice of Public Health

(EPHPP). The QAT checklist evaluates six domains: selection bias, study design, control for confounding, data collection, blinding, and management of dropouts or withdrawals. Each domain was rated as 1 = strong, 2 = moderate, 3 = weak, or NA = not applicable. A summary of scores is provided in Table 1.

7. Synthesis of Results

The authors analyzed data, identifying discrepancies and ensuring consistency across included studies.

III. Results

1. Risk of Bias within Studies

The EPHPP QAT checklist was applied to evaluate the quality of the selected studies. All studies met the objectives of this review and achieved strong ratings across the six evaluation categories.

2. Study Characteristics

A total of 16 studies published between 2020 and 2024 were included, as noted previously. All articles addressed large-

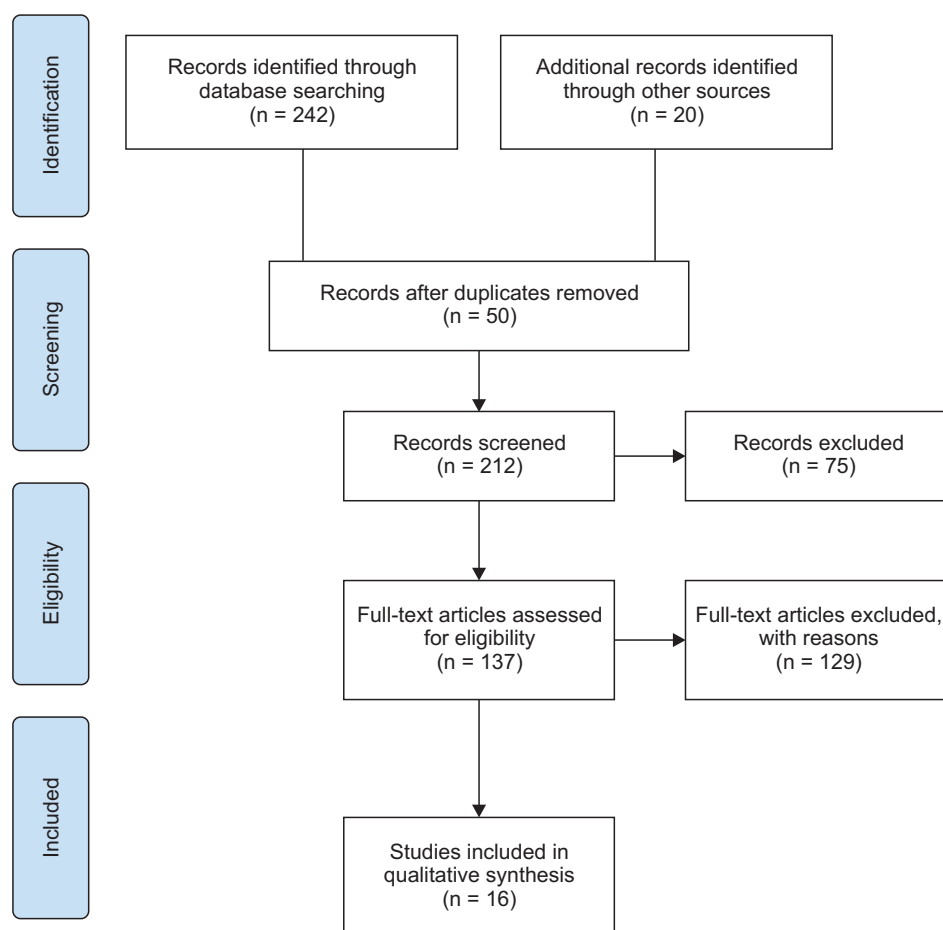


Figure 1. Preferred Reporting Items for Systematic Reviews and Meta-Analyses flow chart.

Table 1. Effective Public Health Practice Project (EPHPP) quality assessment (n = 16)

Study	Selection bias	Study design	Control of confounders	Blinding	Data collection method	Withdrawal and dropout	Analysis
Leopold et al. [40]	NA	1	NA	NA	1	NA	1
Krittanawong et al. [42]	NA	1	NA	NA	1	NA	1
Li et al. [24]	1	1	1	NA	1	NA	1
Adedinsewo et al. [48]	1	1	1	NA	1	NA	1
Kresoja et al. [41]	NA	1	NA	NA	1	NA	1
Hani & Ahmad [13]	1	2	1	2	1	2	1
Hani et al. [4]	1	1	1	NA	1	1	1
Hani et al. [25]	1	2	1	1	1	1	1
Dai et al. [45]	NA	1	NA	NA	1	NA	1
Huang et al. [32]	NA	1	NA	NA	1	NA	1
Subramani et al. [27]	1	1	1	NA	1	NA	1
Muthulakshmi & Parveen [26]	1	1	NA	NA	1	NA	1
Khdaïr [31]	1	1	NA	1	1	NA	1
Hossain et al. [28]	1	1	1	1	1	NA	1
Liu et al. [29]	1	1	NA	1	1	NA	1
Dorraki et al. [30]	1	1	1	1	1	1	1

Each domain was rated as 1 = strong, 2 = moderate, 3 = weak, or NA = not applicable.

scale data or big data used to predict CVD through ML and AI. Four articles were published in 2023 (22.2%), three in 2022 (16.6%), and two each in 2020, 2021, and 2022 (11.1% per year). Five studies were published in 2024 and 2025 (3.6%). A summary of study characteristics is presented in Table 2.

3. Results of Individual Studies

1) Large-scale data analysis techniques and ML algorithms

Big data generated from sources such as the internet, mobile applications, social networks, biomedical records, and patient health data represents a vast and complex volume of information that is difficult to manage using traditional methods. Novel analytic techniques emphasize velocity, veracity, and volume.

Li et al. [24] evaluated the variance between statistical and ML methods in predicting CVD risk at both individual and population levels. Their results showed significant variation in model performance, particularly among high-risk patients. Reported differences between QRISK3 and neural networks ranged from -23.2% to 0.1%, underscoring the need for further research. Similarly, studies by Hani and Ahmad [13,25] predicting IHD in men and women demonstrated that the Chi-squared Automatic Interaction Detectin

(CHAID) model improved accuracy and led to a higher AUC.

Similarly, a study performed by Muthulakshmi and Parveen [26] employed experimental methods to identify relevant predictive features for CVD. Their evaluation showed that the Regularized Principal Components and Quadratic Weighted Entropy Boosting (RPC-QWEB) approach consistently outperformed traditional learning methods across multiple performance measures. Compared with benchmark approaches, RPC-QWEB achieved gains of 3% in accuracy and 5% in sensitivity, along with reductions of 7% in prediction time and 29% in error rate.

2) ML algorithms and AI for predicting CVDs

Subramani et al. [27] conducted a study analyzing data from internet of things devices. Their research demonstrated that artificial neural network structures, when trained on data from multiple medical institutions, achieved 96% accuracy and robust results across several metrics.

Hossain et al. [28] addressed CVD prediction in a lower-middle-income country (LMIC), where CVD is a growing cause of mortality. Their work highlighted the importance of context-specific datasets for improving generalizability across diverse populations. By comparing multiple classi-

Table 2. Matrix of included studies (n = 16)

Study, year	Study design	Country	Sample size	Objective (s)	Main results	Conclusion
Leopold et al. [40], 2020	Review/Classical and contemporary models	USA	Meta-analysis of 184,305 individuals with MI and referents	To demonstrate the genotypic foundations of (endo) phenotypic diversity. To draw attention to the current difficulties in the genotyping and phenotyping of cardiovascular disease, which can be resolved by integrating big data and interpreting the results using cutting-edge analytical techniques (network analysis).	Only when big data collection is subjected to special, integrated analysis techniques that embrace genetic and clinical heterogeneity rather than ignore or diminish it will precision cardiovascular treatment be widely applied to CVD patients.	Big data in conjunction with cutting-edge analytical techniques, such as network analysis, will be able to clarify the genotype–endophenotype–cardiovascular disease link as promoted by network medicine and provide light on the causes of heterogeneity in cardiovascular illnesses.
Krittananawong et al. [42], 2023	Review	USA	Not mentioned	Describe the current state of cardiovascular monitoring from the perspective of biosignal capture, new biosensor discovery, analytical technique development, and, finally, ethical and regulatory concerns.	The cardiac care provided to ambulatory patients could be completely transformed by this expanded monitoring paradigm once data security and other ethical and legal issues about wearable devices are resolved.	The cardiac care provided to ambulatory patients could be completely transformed by this expanded monitoring paradigm once data security and other ethical and legal issues about wearable devices are resolved.
Li et al. [24], 2020	Longitudinal cohort study	England	391 general practices in England enrolled 3.6 million patients from the Clinical Practice Research Datalink.	To evaluate how well statistical and ML methods forecast cardiovascular disease risks at the individual and population levels, as well as how filtering affects risk estimates.	The performance of different models for cardiovascular disease prediction varied significantly, especially in patients at higher risk. The risk variations between QRISK3 and neural networks varied from -23.2% to 0.1%, highlighting the need for further research.	Different models for the same patients' risks significantly differ, suggesting the use of survival models like QRISK3, which considers censoring, and regular consistency assessment as crucial for clinical decision-making.

Continued on the next page.

Table 2. Continued

Study, year	Study design	Country	Sample size	Objective (s)	Main results	Conclusion
Adedinsewo et al. [48], 2022	Screening for all ages of women	USA (Mayo-Clinic)	Not mentioned	This summary provides a comprehensive overview of the potential applications of artificial intelligence and digital technologies in cardiovascular screening in women.	Women are disproportionately affected by CVD and are less likely to receive a diagnosis or treatment. This presents a great opportunity to use AI and digital tools to change the way women receive cardiovascular care, starting with screening.	Even though AI has the potential to be useful in CVD screening, further research is required to determine whether digital technologies live up to the hype enhance patient outcomes, and eliminate disparities. Additionally, the creation and deployment of AI-powered communication and decision support tools integrated into current EHR systems are expensive, and before widespread use, the cost-effectiveness of such tactics must be assessed.
Kresoja et al. [41], 2023	Review	-	Not mentioned	To explain fundamental ideas as well as the risks associated with applying these techniques, aiming to equip contemporary doctors and researchers for the challenges that ML may present. Additionally, a synopsis of the state-of-the-art established classical and developing principles of ML disease prediction in the domains of basic science, imaging, and omics is provided.	Traditional statistical methods may be more useful than cutting-edge ML techniques. To test algorithms in future clinical trials, given the intricate interaction of factors employed in ML, which goes beyond what the human brain can handle. To accept the results of the ML algorithm, even if we are unable to understand every calculation it performs.	Although ML offers many opportunities, there are drawbacks as well. Is crucial to remember that ML approaches are not always the best option, particularly in basic research where causal relationships between variables may be more significant than optimized predictions.

Continued on the next page.

Table 2. Continued

Study, year	Study design	Country	Sample size	Objective (s)	Main results	Conclusion
Hani & Ahmad [13], 2024	Retrospective, predictive study	Jordan	A total of 3,435 EHR of Jordanian males were extracted.	The Chi-squared Automatic Interaction Detection (CHAID) model should be suggested as a prediction method to estimate the likelihood of death versus life in men who may suffer a heart attack.	The CHAID algorithm has a greater accuracy of 93.72% and an area under the curve of 0.792, making it the best top model devised to predict mortality among Jordanian men. Governorates, age, pulse oximetry, medical diagnosis, pulse pressure, heart rate, systolic blood pressure, and pulse pressure were found to be the most important predicted risk variables of heart attack mortality among Jordanian men.	Demographic traits and hemodynamic readings were provided, with heart attack complaints serving as the main risk factors that were predicted using ML methods such as the CHAID model.
Hani et al. [4], 2023	Systematic review	Jordan	13 published articles were eligible for inclusion.	To list and assess the best ML method for ischemic heart disease prediction	3 main topics emerged from the data: the most widely used algorithm for IHD prediction, the precision of ischemic heart disease prediction algorithms, and the clinical outcomes that enhance care quality. Both supervised and unsupervised ML have been used in all approaches.	ML is expected to aid clinicians in interpreting patient data, implementing optimal algorithms, and building evidence-based management for invasive procedures like catheterizations.

Continued on the next page.

Table 2. Continued

Study, year	Study design	Country	Sample size	Objective (s)	Main results	Conclusion
Hani et al. [25], 2023	Retrospective design	Jordan	Big data retrieved from EHR of 2028 women with heart disease	To predict mortality caused by heart disease among women, using AI-based ML algorithm.	The CHAID model achieved the best accuracy (93.25%) and AUC (0.825). The primary predictors of death were age, systolic blood pressure readings with a cutoff value of >187 mm Hg, medical diagnosis (women with congestive heart failure were at a higher risk of death [n=31, 16.58%]), pulse pressure with a cutoff value of 98 mmHg, and oxygen saturation (measured using pulse oximetry) with a cutoff value of 93%.	A useful tool for clinical prediction, the CHAID model (ML algorithm) validated the accurate identification of the major determinants of cardiovascular death among women.
Dai et al. [45], 2022	Review	-	A set of 229 patients suffering from AHF vs. a set of 201 patients with CHF.	To demonstrate the potential benefits of using “Big Data” in cardiology, particularly in the fight against and mitigation of the burden of CVD.	Big Data has the potential to completely transform clinical research and cardiology practice. A few methodological concerns should be appropriately handled, and a few ethical concerns should be taken into account.	Big Data is becoming relevant in today’s world. It permeates every aspect of modern civilization and opens up previously unimaginable avenues for biomedicine, including cardiology.

Continued on the next page.

Table 2. Continued

Study, year	Study design	Country	Sample size	Objective (s)	Main results	Conclusion
Huang et al. [32], 2022	Review	-	-	To cite and explain current advancements in the use of digital health for CVD, with an emphasis on AI methods for wearable data-driven model creation, diagnosis, and prognosis of CVD.	Traditional ML techniques, which rely on hand-engineered features, have been used in the examined literature for ECG data processing and modeling. These techniques include supervised, unsupervised, ensemble, and other rule-based or statistical approaches. While many features are generated, feature selection techniques are mostly required, and domain knowledge is necessary for effective feature extraction.	With the quick advancement of computing and telecommunications technology, digital health-based healthcare solutions have been integrated into cardiology research and on-site practice. This has the potential to enhance current clinical care by empowering patients, both dominant and recessive, to participate more actively in their care.
Subramani et al. [27], 2023	Data observation for monitoring data and methods for training various algorithms	-	918 unique sample data-sets. Started with 1,190 samples	To analyze and make predictions based on the data that IoT devices receive.	Deep learning research can benefit from more data from multiple medical institutes because it could be used to build artificial neural network structures. To confirm the effectiveness of, integrated the Heart Dataset with additional classification models. Compared to other current methods, the suggested method yields accuracy results of about 96%, and a thorough study across multiple metrics has been provided.	Novel risk factors for CVDs as well as new genotypes and phenotypes for a range of CVDs may be found through ML. Applying more advanced ML techniques, such as deep learning and artificial neural networks, can improve every element of medical picture recognition, diagnosis, outcome prediction, and prognosis evaluation.

Continued on the next page.

Table 2. Continued

Study, year	Study design	Country	Sample size	Objective (s)	Main results	Conclusion
Muthulakshmi & Parveen [26], 2024	Experimental evaluation based on accuracy, precision, time, and sensitivity of CVD prediction.		Regularized Principal Components and Quadratic Weighted Entropy Boosting (RPC-QWEB) for predicting heart disease	To identify relevant features utilizing ML techniques for predicting cardiovascular disease.	The RPC-QWEB approach performs better than traditional learning approaches over a wide range of performance matrices. Compared to the current benchmark approaches, the RPC-QWEB method yields a 3% and 5% gain in accuracy and sensitivity as well as a 7% and 29% reduction in prediction time and error rate.	The simulation results show that, in terms of many performance matrices, the suggested RPC-QWEB method performs better than the traditional learning methods. The goal of the work going forward is to efficiently forecast cardiac disease using ML and deep learning techniques that will yield better results in less time. Other performance measures like the F1 Score, false positive rate, and others might also be taken into account.
Khdaïr [31], 2021	Comparative analysis		462 medical instances and 9 features as well as the class feature from the South African Heart Disease data retrieved from the KEEL repository. The dataset consists of 302 records of healthy patients and 160 records of patients who suffer from CHD	To compare various ML methods that can reliably forecast the occurrence of cardiac arrest episodes based on clinical data.	Support vector machine has the best overall prediction performance, according to the findings presented using various evaluation metrics.	ML success relies on data richness, and while the selected dataset has known features and risk factors for predicting CHD, more data and variables could improve prediction results. Additional external datasets could validate findings.

Continued on the next page.

Table 2. Continued

Study, year	Study design	Country	Sample size	Objective (s)	Main results	Conclusion
Hossain et al. [28], 2024	Large datasets were extracted from the National Institute of Cardiovascular Disease		A random sample of 391 CVD patients' medical records	To accurately predict the prevalence and risk factors of cardiac disease in Bangladesh	The Random Forest classifier achieved the highest accuracy with 98.04%, achieving the highest precision, robust recall, and high F1 score, surpassing the Logistic Regression model's lowest accuracy.	The study identifies critical factors impacting CVD patients and predicts risk, recommending the Random Forest technique for cardiac disease prediction, potentially enhancing clinical practice and patient prognosis.
Liu et al. [29], 2025	Systematic review and meta-analysis		32 ML models and 26 conventional statistical models from 20 selected studies, focusing on performance metrics such as AUC and heterogeneity across models	To evaluate and compare the efficacy of ML models against conventional CVD risk prediction algorithms using EHR data for medium to long-term (5–10 years) CVD risk prediction	ML models for CVD risk have improved calibration but face variability and methodological concerns, requiring further research to improve transparency and standardization.	The study emphasizes the superiority of ML models in predicting CVD risk and underscores the necessity for rigorous validation for their integration into clinical practice.
Dorraki et al. [30], 2024	A correlational study	UK	A comprehensive UK Biobank dataset (n = 375,145) was used.	To assess the impact of integrating psychological data into a novel ML approach on enhancing CVD prediction performance.	The ensemble ML model outperformed all five constituent learning algorithms in predicting CVD with 71.31% accuracy using traditional risk factors and 85.13% accuracy when psychological factors were added.	The integration of mental health assessment data into an ensemble ML model leads to a highly accurate and improved CVD prediction model.

MI: myocardial infarction, CVD: cardiovascular disease, ML: machine learning, AI: artificial intelligence, EHR: electronic health record, IHD: ischemic heart disease, CHD: coronary artery disease, AUC: area under the curve.

fiers, they provided a comprehensive evaluation of model performance, showing that ensemble approaches such as random forest can enhance accuracy by combining weak learners.

A meta-analysis study by Liu et al. [29] reviewed 22 studies published between 2010 and 2024 that used ML models with EHR data for predicting CVD risk over 5–10 years. They compared algorithms such as gradient boosting machine, support vector machines (SVM), and convolutional neural networks against traditional risk scores like QRISK3 and ASCVD. Their results showed that ML consistently outperformed conventional scores, with gradient boosting machine achieving a pooled AUC of 0.89 (95% confidence interval [CI], 0.85–0.92). Another study performed by Dorraki et al. [30] analyzed data from the UK Biobank ($n = 375,145$) to examine the contribution of psychological variables, such as mental health disorders, in ML-based CVD prediction. Their ensemble approach combined decision tree, random forest, extreme gradient boosting (XGBoost), SVM, and deep neural networks, tested both with and without psychological factors. The inclusion of psychological data improved the AUC from 0.82 to 0.87, with XGBoost demonstrating the best performance. This novel focus on mental health addresses a significant gap in conventional CVD risk models, which typically exclude psychological determinants despite their bidirectional relationship with CVD. The large sample size and use of the UK Biobank increased statistical power and generalizability, while the ensemble strategy leveraged the strengths of multiple algorithms to enhance predictive accuracy.

Khdaier [31] conducted a comparative analysis evaluating various ML methods for forecasting cardiac arrest episodes based on clinical data. The study found that SVMs provided the best overall predictive performance, as determined by multiple evaluation metrics.

Similarly, Huang et al. [32] highlighted current advancements in the use of digital health for CVD, emphasizing AI applications in wearable data-driven model development, diagnosis, and prognosis. Their study underscored the growing role of wearable technologies in generating high-frequency, real-world data to improve cardiovascular risk assessment.

In another study, Muthulakshmi and Parveen [26] identified relevant features for CVD prediction using ML techniques. Their RPC-QWEB approach outperformed conventional learning methods across a wide range of performance metrics. Compared with benchmark models, RPC-QWEB achieved 3% and 5% improvements in accuracy and sensitivity, along with reductions of 7% in prediction time and 29%

in error rate. Simulation results confirmed that the method outperforms traditional learning approaches across multiple evaluation matrices. The authors emphasized that future work should focus on optimizing ML and deep learning strategies to achieve faster and more accurate cardiac disease prediction. They also recommended expanding performance assessment to include additional measures such as the F1-score and false positive rate.

Across the reviewed literature, CVD prediction models demonstrated diverse clinical objectives. The prediction of major adverse cardiac events was the most commonly targeted endpoint, particularly in asymptomatic populations [33]. Other models, often incorporating imaging or biomarker data, focused on outcomes such as hospitalization for heart failure, onset of atrial fibrillation, or progression of coronary artery disease [34,35]. More advanced models sought to predict short-term outcomes, including 30-day readmission or in-hospital mortality following acute coronary syndrome [36]. A smaller but growing body of research has targeted cardiovascular-specific and all-cause mortality, using longitudinal data to stratify long-term risk. Clearly defining these outcomes is essential, as the endpoint directly determines a model's clinical utility, whether in long-term management, acute event prediction, or early screening.

To conclude, models predict IHD using large-scale data and ML displayed varying levels of accuracy, sensitivity, and specificity. Their strengths and limitations depend largely on the type of data employed and the specific clinical context.

3) Role of integrating large-scale data in disease prediction to improve quality of care

Healthcare organizations increasingly employ large-scale data analytics and advanced technologies to enhance clinical outcomes and reduce costs [37]. ML is becoming embedded in daily medical practice and holds the potential to significantly transform modern medicine [38].

Large-scale data encompasses enormous and continuously expanding information streams, including digital signals, text, images, and videos [39]. Analyzing such data requires high computational capacity and advanced analytical techniques such as component analysis, latent semantic analysis, and dimensionality reduction.

A systematic review conducted by Leopold et al. [40] recommended leveraging big data with advanced methods such as network analysis to elucidate genotype–endophenotype–CVD relationships, as advocated in network medicine. This integration could help explain heterogeneity in cardiovascular disease presentation and outcomes. Conversely, Kresoja

et al. [41] aimed to explain fundamental ideas as well as the risks associated with applying these techniques, emphasizing the need to equip physicians and researchers with a clear understanding of ML's limitations. Their study reported that although ML provides substantial opportunities for predicting CVD outcomes, it also introduces challenges and potential drawbacks that must be carefully managed.

IV. Discussion

The studies included in this review highlighted the role of large-scale data in predicting CVD and improving healthcare outcomes. Four studies demonstrated the critical contribution of ML and AI in predicting CVD using big data and in building evidence-based management strategies for invasive procedures such as catheterization [31,40-42]. Similarly, four studies identified the CHAID algorithm and SVM as among the most accurate models for predicting CVD, yielding the highest accuracy and AUC across different target populations [13,25,31].

Analytics leverage real-world data from EHR, insurance claims, and wearable devices to assess treatment outcomes and comparative effectiveness in diverse CVD populations. For example, causal inference models employing propensity score matching adjust for confounding factors in observational data, thereby generating insights into treatment effects beyond controlled clinical trials [43]. In addition, Armoundas et al. [44] emphasized that developing tractable, cost-effective AI- and ML-based precision medicine requires the establishment of implementation science frameworks. These frameworks must embed strategies to minimize bias and enhance generalizability to ensure that AI/ML tools do not perpetuate existing healthcare disparities but instead generate evidence that can be validated in clinical trials.

Three studies demonstrated the range of technological approaches available for predicting CVD, including simulation results for the RPC-QWEB method, which consistently outperforms traditional learning approaches. Future research should aim to forecast cardiac disease efficiently using ML and deep learning techniques that improve predictive performance while reducing processing time. Broader evaluation metrics, such as the F1-score and false positive rate, should also be incorporated [24,32,45]. Hossain et al. [28] underscored the importance of developing ML models for resource-limited settings, thereby addressing gaps in Western-centric CVD literature. While their work highlighted the adaptability of ML to local datasets, it also revealed a lack of interpretability analyses tailored to clinical needs. Similarly,

Liu et al. [29] through their meta-analysis, demonstrated that ML consistently outperforms traditional risk scores in clinical applications but noted significant barriers to clinical adoption and stressed the need for continued systematic reviews. Dorraiki et al. [30] advanced the field further by integrating mental health variables into predictive models, offering a more holistic approach to CVD prediction. Their findings align with precision medicine goals by tailoring risk assessment to individual psychological profiles, though validation across diverse populations remains necessary for broader applicability.

Some studies emphasized that while ML presents many opportunities, it also has notable limitations. ML approaches may not always be optimal, particularly in basic research contexts where identifying causal relationships is more valuable than maximizing predictive accuracy.

AI and ML are increasingly improving CVD research by analyzing EHR, cardiac imaging, and genetic datasets, yet challenges such as high-dimensional signals and imbalanced data remain [46]. Despite these barriers, ML algorithms have significantly improved the accuracy of automated diagnosis in cardiac imaging, and AI-driven risk prediction systems have shown promise in patient stratification for early intervention [13,47]. Additionally, wearable and remote monitoring devices provide continuous, individualized cardiovascular data, enabling proactive patient management [48]. Ongoing research seeks to refine these tools by enhancing transparency, generalizability, and clinical utility in real-world settings [4].

Large-scale data analytics contributes to CVD research by uncovering insights from complex datasets and supporting clinical practice through personalized, real-time, predictive tools. At the same time, it raises challenges related to ethics, technical implementation, and healthcare equity.

The integration of big data improves risk identification, facilitates early intervention, and enables personalized treatment strategies. ML algorithms detect patterns in patient data, while large-scale data integration enhances decision-making through wearables and telemedicine platforms, supporting remote monitoring and early detection of adverse events [49].

CVD development is shaped by interactions among socioeconomic factors, lifestyle, environment, and genetics. Large-scale data analytics continues to struggle in fully capturing these multidimensional relationships, making the creation of a universal CVD detection model difficult due to population-specific differences in risk factors.

Overall, this review examined the use of large-scale data

analysis techniques and ML algorithms for predicting CVD, highlighting the value of integrating these technologies to improve healthcare quality. Nevertheless, issues such as bias, data quality, ethical implications, and practical deployment demand cross-disciplinary collaboration, continuous refinement, and further research.

Conflict of Interest

No potential conflict of interest relevant to this article was reported.

ORCID

Salam Bani Hani (<https://orcid.org/0000-0003-0848-5615>)

Muayyad M. Ahmad (<https://orcid.org/0000-0002-4388-8332>)

References

1. World Health Organization. Cardiovascular diseases (CVDs) [Internet]. Geneva, Switzerland: World Health Organization; 2025 [cited at 2025 Oct 13]. Available from: [https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-\(cvds\)](https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)).
2. Heron M. Deaths: leading causes for 2019. *Natl Vital Stat Rep* 2021;70(9):1-114.
3. Mc Namara K, Alzubaidi H, Jackson JK. Cardiovascular disease as a leading cause of death: how are pharmacists getting involved? *Integr Pharm Res Pract* 2019;8:1-11. <https://doi.org/10.2147/IPRPS133088>
4. Hani SB, Ahmad MM. Machine-learning algorithms for ischemic heart disease prediction: a systematic review. *Curr Cardiol Rev* 2023;19(1):e090622205797. <https://doi.org/10.2174/1573403X18666220609123053>
5. Sharma P, Sharma P, Sharma K, Varma V, Patel V, Sarvaiya J, et al. Revolutionizing utility of big data analytics in personalized cardiovascular healthcare. *Bioengineering (Basel)* 2025;12(5):463. <https://doi.org/10.3390/bioengineering12050463>
6. Madaksira U. Risk assessment of cardiovascular diseases utilizing individual lifestyle factors with machine learning techniques [master's thesis]. Dublin, Ireland: Dublin Business School; 2024.
7. Jamthikar A, Gupta D, Cuadrado-Godia E, Puvvula A, Khanna NN, Saba L, et al. Ultrasound-based stroke/ cardiovascular risk stratification using Framingham Risk Score and ASCVD Risk Score based on "Integrated Vascular Age" instead of "Chronological Age": a multi-ethnic study of Asian Indian, Caucasian, and Japanese cohorts. *Cardiovasc Diagn Ther* 2020;10(4):939-54. <https://doi.org/10.21037/cdt.2020.01.16>
8. Traina MI, Almahmeed W, Edris A, Murat Tuzcu E. Coronary heart disease in the Middle East and North Africa: current status and future goals. *Curr Atheroscler Rep* 2017;19(5):24. <https://doi.org/10.1007/s11883-017-0659-9>
9. Ahmad M, Hani SB. Differences in ischemic heart disease between males and females using predictive artificial intelligence models. *Electr J Gen Med* 2024;21(5):em607. <https://doi.org/10.29333/ejgm/15174>
10. Nachiappan S, Chandran A, Swathika B, Ganesan S, Mahaprasad A, Muddebihal F, et al. Pulp stones: diagnostic significance in early diagnosis and radiographic correlation with ischemic heart diseases. *Indian J Radiol Imaging* 2021;31(2):277-83. <https://doi.org/10.1055/s-0041-1731829>
11. Del Re DP. Mechanisms of ischemic heart injury. *Cells* 2022;11(9):1384. <https://doi.org/10.3390/cells11091384>
12. Albus C, Barkhausen J, Fleck E, Haasenritter J, Lindner O, Silber S. The diagnosis of chronic coronary heart disease. *Dtsch Arztebl Int* 2017;114(42):712-9. <https://doi.org/10.3238/arztebl.2017.0712>
13. Hani SB, Ahmad M. Predicting mortality amongst Jordanian men with heart attacks using the chi-square automatic interaction detection model. *Health Inform J* 2024;30(3):14604582241270830. <https://doi.org/10.1177/14604582241270830>
14. Kondraganti A, Narayanamurthy G, Sharifi H. A systematic literature review on the use of big data analytics in humanitarian and disaster operations. *Ann Oper Res* 2024;335:1015-52. <https://doi.org/10.1007/s10479-022-04904-z>
15. Theodorakopoulos L, Theodoropoulou A, Stamatiou Y. A state-of-the-art review in big data management engineering: real-life case studies, challenges, and future research directions. *Eng* 2024;5(3):1266-97. <https://doi.org/10.3390/eng5030068>
16. Nwaimo CS, Adegbola AE, Adegbola MD. Transforming healthcare with data analytics: Predictive models for patient outcomes. *GSC Biological and Pharmaceutical Sciences* 2024;27(3):25-35. <https://doi.org/10.30574/gscbps.2024.27.3.0190>
17. Dykstra S. Integrating multi-domain electronic health data, machine learning, and automated cardiac phenomics for personalized cardiovascular care [master's thesis]. Calgary, Canada: University of Calgary; 2024

- <https://doi.org/10.11575/PRISM/43304>
18. Rydin AO, Aalbers G, van Eeden WA, Lamers F, Milan-eschi Y, Penninx BW. Predicting incident cardio-metabolic disease among persons with and without depressive and anxiety disorders: a machine learning approach. *Soc Psychiatry Psychiatr Epidemiol* 2025;60(6):1457-66. <https://doi.org/10.1007/s00127-025-02857-9>
 19. Meng C, Trinh L, Xu N, Enouen J, Liu Y. Interpretability and fairness evaluation of deep learning models on MIMIC-IV dataset. *Sci Rep* 2022;12(1):7166. <https://doi.org/10.1038/s41598-022-11012-2>
 20. Ahmad M, Hani SH, Sabra MA, Almahmoud O. Big data can help prepare nurses and improve patient outcomes by improving quality, safety, and outcomes. *Front Nurs* 2023;2:241-8. <https://doi.org/10.2478/FON-2023-0026>
 21. Singhanian K, Reddy A. Improving preventative care and health outcomes for patients with chronic diseases using big data-driven insights and predictive modeling. *Int J Appl Heal Care Anal* 2024;9(2):1-14.
 22. Dang D, Dearholt SL, Bissett K, Ascenzi J, Whalen M. Johns Hopkins evidence-based practice for nurses and healthcare professionals: model and guidelines. 4th ed. Indianapolis (IN): Sigma Theta Tau International; 2021.
 23. Thomas BH, Ciliska D, Dobbins M, Micucci S. A process for systematically reviewing the literature: providing the research evidence for public health nursing interventions. *Worldviews Evid Based Nurs* 2004;1(3):176-84. <https://doi.org/10.1111/j.1524-475X.2004.04006.x>
 24. Li Y, Sperrin M, Ashcroft DM, van Staa TP. Consistency of variety of machine learning and statistical models in predicting clinical risks of individual patients: longitudinal cohort study using cardiovascular disease as exemplar. *BMJ* 2020;371:m3919. <https://doi.org/10.1136/bmj.m3919>
 25. Hani SB, Ahmad M. Effective Prediction of mortality by heart disease among women in Jordan using the chi-squared automatic interaction detection model: retrospective validation study. *JMIR Cardio* 2023;7:e48795. <https://doi.org/10.2196/48795>
 26. Muthulakshmi P, Parveen M. Big data analytics for heart disease prediction using regularized principal and quadratic entropy boosting. *Indian J Sci Technol* 2024;17(6):533-47.
 27. Subramani S, Varshney N, Anand MV, Soudagar MEM, Al-Keridis LA, Upadhyay TK, et al. Cardiovascular diseases prediction by machine learning incorporation with deep learning. *Front Med (Lausanne)* 2023;10:1150933. <https://doi.org/10.3389/fmed.2023.1150933>
 28. Hossain S, Hasan MK, Faruk MO, Aktar N, Hossain R, Hossain K. Machine learning approach for predicting cardiovascular disease in Bangladesh: evidence from a cross-sectional study in 2023. *BMC Cardiovasc Disord* 2024;24(1):214. <https://doi.org/10.1186/s12872-024-03883-2>
 29. Liu T, Krentz A, Lu L, Curcin V. Machine learning based prediction models for cardiovascular disease risk using electronic health records data: systematic review and meta-analysis. *Eur Heart J Digit Health* 2025;6(1):7-22. <https://doi.org/10.1093/ehjdh/ztae080>
 30. Dorraki M, Liao Z, Abbott D, Psaltis PJ, Baker E, Bidargaddi N, et al. Improving cardiovascular disease prediction with machine learning using mental health data: a prospective UK Biobank study. *JACC Adv* 2024;3(9):101180. <https://doi.org/10.1016/j.jacadv.2024.101180>
 31. Khadair H. Exploring machine learning techniques for coronary heart disease prediction. *Int J Adv Comput Sci Appl* 2021;12(5):28-36.
 32. Huang JD, Wang J, Ramsey E, Leavey G, Chico TJA, Condell J. Applying artificial intelligence to wearable sensor data to diagnose and predict cardiovascular disease: a review. *Sensors (Basel)* 2022;22(20):8002. <https://doi.org/10.3390/s22208002>
 33. Almustafa KM. Prediction of heart disease and classifiers' sensitivity analysis. *BMC Bioinformatics* 2020;21(1):278. <https://doi.org/10.1186/s12859-020-03626-y>
 34. Pocock SJ, Ferreira JP, Gregson J, Anker SD, Butler J, Filippatos G, et al. Novel biomarker-driven prognostic models to predict morbidity and mortality in chronic heart failure: the EMPEROR-Reduced trial. *Eur Heart J* 2021;42(43):4455-64. <https://doi.org/10.1093/eurheartj/ehab579>
 35. Nowak C, Arnlov J. Kidney disease biomarkers improve heart failure risk prediction in the general population. *Circ Heart Fail* 2020;13(8):e006904. <https://doi.org/10.1161/CIRCHEARTFAILURE.120.006904>
 36. Cholack G, Garfein J, Krallman R, Feldeisen D, Montgomery D, Kline-Rogers E, et al. Predictors of early (0-7 days) and late (8-30 days) readmission in a cohort of acute coronary syndrome patients. *Int J Med Stud* 2022;10(1):38-48. <https://doi.org/10.5195/ijms.2022.1058>
 37. Dash S, Shakyawar SK, Sharma M, Kaushik S. Big data in healthcare: management, analysis and future prospects. *J Big Data* 2019;6(1):54. <https://doi.org/10.1186/s40537-019-0217-0>

38. Oakden-Rayner L. Exploring large-scale public medical image datasets. *Acad Radiol* 2020;27(1):106-12. <https://doi.org/10.1016/j.acra.2019.10.006>
39. Sarwar Kamal M, Dey N, Ashour AS. Large scale medical data mining for accurate diagnosis: a blueprint. In: Khan S, Zomaya A, Abbas A, editors. *Handbook of large-scale distributed computing in smart healthcare*. Cham, Switzerland: Springer; 2017. p. 157-76. https://doi.org/10.1007/978-3-319-58280-1_7
40. Leopold JA, Maron BA, Loscalzo J. The application of big data to cardiovascular disease: paths to precision medicine. *J Clin Invest* 2020;130(1):29-38. <https://doi.org/10.1172/JCI129203>
41. Kresoja KP, Unterhuber M, Wachter R, Thiele H, Lurz P. A cardiologist's guide to machine learning in cardiovascular disease prognosis prediction. *Basic Res Cardiol* 2023;118(1):10. <https://doi.org/10.1007/s00395-023-00982-7>
42. Krittanawong C, Rogers AJ, Johnson KW, Wang Z, Turakhia MP, Halperin JL, et al. Integration of novel monitoring devices with machine learning technology for scalable cardiovascular management. *Nat Rev Cardiol* 2021;18(2):75-91. <https://doi.org/10.1038/s41569-020-00445-9>
43. Dhingra LS, Shen M, Mangla A, Khera R. Cardiovascular care innovation through data-driven discoveries in the electronic health record. *Am J Cardiol* 2023;203:136-48. <https://doi.org/10.1016/j.amjcard.2023.06.104>
44. Armoundas AA, Narayan SM, Arnett DK, Spector-Bagdady K, Bennett DA, Celi LA, et al. Use of artificial intelligence in improving outcomes in heart disease: a scientific statement from the American Heart Association. *Circulation* 2024;149(14):e1028-50. <https://doi.org/10.1161/CIR.0000000000001201>
45. Dai H, Younis A, Kong JD, Puce L, Jabbour G, Yuan H, et al. Big data in cardiology: state-of-art and future prospects. *Front Cardiovasc Med* 2022;9:844296. <https://doi.org/10.3389/fcvm.2022.844296>
46. Sultana S, Hriday MS, Haque R, Das M. Health monitoring through wearables: a systematic review of innovations in cardiovascular disease detection and prevention. *Strateg Data Manag Innov* 2025;2(1):96-115.
47. Hani SB, Ahmad M. Using big data to predict young adult ischemic vs. non-ischemic heart disease risk factors: an artificial intelligence based model. *Intell Based Med* 2025;11:100207. <https://doi.org/10.1016/j.ibmed.2025.100207>
48. Adedinsewo DA, Pollak AW, Phillips SD, Smith TL, Svatikova A, Hayes SN, et al. Cardiovascular disease screening in women: leveraging artificial intelligence and digital tools. *Circ Res* 2022;130(4):673-90. <https://doi.org/10.1161/CIRCRESAHA.121.319876>
49. Li Y, Sperrin M, Martin GP, Ashcroft DM, van Staa TP. Examining the impact of data quality and completeness of electronic health records on predictions of patients' risks of cardiovascular disease. *Int J Med Inform* 2020;133:104033. <https://doi.org/10.1016/j.ijmedinf.2019.104033>