Transformer-based Models for Cardiovascular Disease Predictions from Electronic Health Records: A Systematic Review

Onayi T Chikumo [1]*, Belinda Ndlovu [2]*
* Informatics and Analytics Department, National University of Science and Technology, Bulawayo, Zimbabwe
n02220001j@students.nust.ac.zw [1], belinda.ndlovu@nust.ac.zw [2]

Article Info

Article history:

Received 2025-11-30
Revised 2026-01-07
Accepted 2026-01-17

Keyword:

Cardiovascular Diseases,
Electronic Health Records,
Explainable AI
Transformer Model,
Multimodal Data,

ABSTRACT

This systematic literature review (SLR) analyses 16 studies published between 2020 and 2025 that applied transformer-based or other machine learning models to predict cardiovascular disease (CVD) using electronic health records (EHRs). Following the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines, the review ensures transparency in the identification, screening, and quality appraisal of eligible studies. The key findings reveal a rapid shift from traditional machine learning models, such as Random Forest, toward transformer architectures like the Bidirectional Encoder Representation from Transformers for Electronic Health Record (BEHRT) and its variants. These models demonstrate a superior discrimination (Area Under Curve:0.84 to 0.93) due to their capacity to model long-term temporal dependencies. Explainable AI (XAI) tools, such as attention visualisation, were frequently employed, yet clinical interpretability and integration into decision support remain underexplored. The review also highlights opportunities in federated and privacy-preserving learning, multimodal data fusion, and hybrid architectures that integrate transformers with traditional machine learning methods. This review addresses a gap in the past literature by being the first SLR to compare transformer variants for the prediction of CVDs. Other SLRs examined general CVD risk models, but the present SLR analyses interpretability, external validation and methodological limitations to transformer models. The findings of the recent SLR reported challenges that include data-shift limitations, model-poor population generalisation and their limitations to clinical adoption, which highlights the need for more evaluation protocols and clinicians' interpretability frameworks.

I. Introduction

Cardiovascular diseases (CVDs) are a group of disorders affecting the heart and blood vessels [1]. CVDs are a leading cause of morbidity and mortality worldwide from Noncommunicable disease [2]-[4], accounting for 17.0 million deaths, representing one-third of global deaths[5], and putting a heavy load on healthcare systems and national economies. CVD mortality is common in the majority of developed, developing, and impoverished nations [1], [6],[7].To address the global burden of CVD [1], healthcare systems are prioritising strategies that predict the onset to enable early intervention [5],[4]. The increased use of Electronic Health Records (EHRs) has created significant opportunities for CVD prediction [8]. EHRs capture a variety of patient data, including demographic information, medical histories and lab results, which offers a longitudinal view of patient data[9] (Figure 1). Traditional risk scores, such as Systematic Coronary Risk Evaluation (SCORE) and QResearch Cardiovascular Risk Algorithm, Version 3 (QRISK3), have been widely used to estimate the risk of CVD; however, their performance is suboptimal in diverse populations [8]. Artificial intelligence(AI) technologies, such as Machine Learning, are used to enhance medical research, personalised treatment and diagnostic accuracy of diseases [10], and when EHRs are integrated with Machine Learning (ML) models, they improve the identification of individuals who are at risk [9]. The researcher developed ML and Deep Learning (DL) predictive models to address traditional methods, demonstrating superior calibration and discrimination compared to conventional strategies [9].

Several studies have applied deep learning techniques to CVD prediction using EHR data. For example, [3] employed a recurrent neural network to model longitudinal patient reports and reported an improved discrimination compared to logistic regression. However, their approach struggled with long-term dependencies and lacked interpretability, and [8] reviewed 79 Artificial Intelligence (AI)-based CVD studies with 486 predictive models that reported high bias despite promising accuracy. Similarly,[11] demonstrated strong predictive performance using a deep neural network but also highlighted limited external validation across the healthcare system. The study [9] revealed that ML models achieved a higher score in predictive accuracy (AUC-0.87) than QRISK3 (AUC-0.77), and reported that the model lacks interpretability and external validation. However, predictive models often face challenges such as institutional bias or missing data and insufficient infrastructure [7], which limits their clinical adoption.

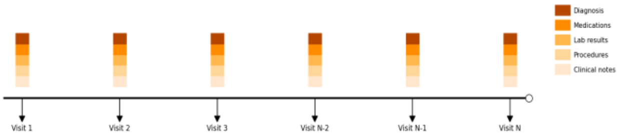

Figure 1. Patient Electronic Health Record

Figure 1 shows an EHR example that includes diagnosis, medication, lab results, procedures, and clinical notes from the date of visit.

More recently, transformer-based architectures have been introduced to address the limitation of sequential models in handling long-term dependencies within EHR data[12]. A researcher in 2020 proposed BEHRT [13], a bidirectional transformer model that represents patient medical histories as sequences of clinical tokens, achieving superior performance in the cardiovascular risk task. Subsequent studies, including Carefully Optimised and Rigorously Evaluated BEHRT (CORE-BEHRT)[13] and Hi-BEHRT [14], refined the original architecture to improve efficiency and interpretability. Similarly [15], implemented prediction on CVD using a transformer-based model (TRisk) which demonstrated a superior performance of (Concordance-index -0.91) compared to QRISK3, later [16] comes with Clinical Electronic Health Records(CEHR)-Bidirectional Encoder Representations from Transformers(BERT), which introduced time token and age embedding into the transformer architecture to help the system identify whether a diagnosis was made recently or over years ago to help in disease predictions.[17] introduced Medical-BERT, which is a transformer-based deep learning model that is designed to work with structured EHRs. The study examined diagnosis codes, medications, and procedures. The Med-BERT learns from diagnosis codes to predict the onset of a risk. The model excludes time-gap embedding that was used in BEHRT [18], which limited temporal precision. [19] introduced an innovation, Multimodal-BEHRT, which integrates textual clinical notes and tabular data for disease predictions. The Multimodal-BEHRT was used to predict Breast cancer.[20] introduced another approach, Targeted-BEHRT, where causal inference capabilities were added to improve interpretability. Despite these advances, existing studies vary widely in evaluation protocols, outcome definitions and validation strategies, making it difficult to draw consistent conclusions regarding clinical applicability.

Although there are growing opportunities in leveraging transformation-based EHR models, no existing systematic literature review (SLR) has examined these models specifically for CVDs prediction, how they validate external, how they compare with traditional ML and their use of explainability tools. Past studies on SLRs have focused on generic models for CVD prediction or on how transformers are used for clinical tasks, leaving a gap in understanding the transformer's specific strengths and weaknesses in CVD modelling using longitudinal EHRs.This SLR addresses these gaps by providing a focused and comparative synthesis of transformer-based models for CVD prediction using EHRs. The study will be guided by the following research questions.

1. What are the most common algorithms and transformer models that have been applied to Cardiovascular disease prediction leveraging EHRS?
2. How do the transformer models perform compared to traditional, deep learning and machine learning approaches based on evaluation metrics such as AUC, F1-score, calibration and accuracy?
3. What explainable artificial intelligence (XAI) techniques have been integrated into Transformer models to improve clinical interpretability?
4. What limitations and opportunities are presented by transformer models for future empirical investigation?

II. Method

This SLR was conducted according to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) [21].

A. *Search Strategy*

A search for relevant studies is conducted across various databases, including PubMed, Xplore Digital Library (IEEE), and Scopus. The search terms used across different databases are summarised as: General search term ("Cardiovascular diseases "OR "Heart disease" OR "CVD") AND ("Electronic Health Records" OR "EHRs") AND ("Prediction" OR "Risk*") AND ("Transformer" OR "BEHRT" OR "Machine learning "OR "Deep learning "OR "Neural Network") that was used across the 3 data sources. Using the search terms, the identified studies were from

PubMed (72), IEEE (5), and Scopus (30), totalling 107 studies from 2020 to 2025 in the English language.

B. *Inclusion and Exclusion Criteria*

The searches were restricted to the following inclusion criteria, as outlined in Table I, and exclusion criteria in Table II, using a Population, Intervention, Comparator, Outcome, Study (PICOS) framework [22].Table I shows the inclusion criteria used in the selection of studies.

TABLE I
INCLUSION CRITERIA

Category	Inclusion	Rationale
Population	Human studies using Electronic Health Records(EHRs) ,>= 18 years related to cardiovascular disease(CVD)	Studies that ensure the clinical relevance of EHR-based models for the prediction task.
Intervention	Studies that utilise transformers like BERT, Hybrid Transformer(Recurrent Neural Network) for prediction, diagnoses or risk assessment of CVD and other traditional Machine Learning.	Studies that target the contribution of transformers or ML models for CVD outcomes.
Comparator	Studies that compare transformer models with other machine learning/deep learning models and use metrics to assess each model.	Allow benchmarking and comparative analysis of model performance.
Outcome	Studies that report quantitative predictive performance metrics like AUC, F1-score and precision.	Ensures the inclusion of studies with measurable outcomes.
Study	Peer-reviewed empirical studies, preprints with methodological details.	Studies that guarantee methodological rigour and reproducibility.
Publication Year	From 2020 -2025, i.e. for the definition of a term, earlier years can be used.	Studies included leveraging current developments for CVD prediction.

Table II presents the exclusion criteria used to guide the selection of studies.

TABLE II
EXCLUSION CRITERIA

Category	Exclusion	Rationale
Data source	Studies not using EHRs	Included are EHR-based models
Outcome	Studies focusing on disease prediction which are not related to CVD	Maintains thematic relevance.
Publication Type	Abstracts that lack a full methodology	Ensures methodological depth
Simulation	Simulated dataset	Highlight real clinical relevance

C. *Screening studies*

The PRISMA flow chart illustrates the 3 databases that were searched, with results of PubMed (72), IEEE (5), and Scopus (30). Mendeley was used to store the downloaded studies and remove duplicate records; the initial total was 107. Records that remained after removing duplicates numbered 97. The screening process consisted of 2 stages: title and abstract screening, and full-text screening. During the abstract and title screening, we assessed the study aim and methods to determine if each paper fell within our scope of review. The number of eliminated studies was 70. A total of 27 papers remained, and they were reviewed to determine whether they addressed the research questions. After a full-text review, 12 documents were excluded. Sixteen studies were included and are listed in Table IV.

D. *Eligibility Criteria*

The studies included were developing a prediction model using EHRs with tabular data, which is structured such as demographic, diagnosis codes, medication and lab tests. The eligible studies had to use ML/DL and transformers to predict the onset of CVDs with model performance evaluation metrics and report performance metrics comparing the baseline models. The studies that integrated any of the Explainable AI or did not integrate were included. The studies that were not eligible were the ones that used other data types (e.g. genomic datasets) to predict the onset of CVDs. The study on a specific population (e.g. studies investigating the performance of predictive models in the HIV-positive population only) was excluded, and those that were using EHRs to predict non-CVDs forecasting. All studies included had to be written in English. Studies that fall between 2020 and 2025 were included to ensure that studies included leveraged current developments for CVD prediction. The researcher [18] introduced a transformer called BEHRT in 2020, which marked the emergence of transformer-based models that reshaped EHR modelling. Restricting the period ensured that the evaluation of modern models and enhanced methodological comparability, given the advancement of AI tools, earlier papers (before 2020) relied most on conventional ML models.

E. Assessment of Study Quality and Risk of Bias

The methodological quality and risk of bias of the included studies were systematically evaluated using the Prediction Model Risk of Bias Assessment Tool(PROBAST). This framework assesses bias across four domains: participants, predictors, outcomes and statistical analysis. Each study was independently assessed by two reviewers (O.C and B.N), with discrepancies resolved through consensus. Each domain was rated as" low risk"," moderate", or "high risk" according to the PROBAST guidelines and the overall risk of bias was assigned to each study. The risk of bias assessments was considered when interpreting the findings. The detailed risk of bias assessment is presented in Table III.Risk of bias assessments were used to contextualise findings across studies, with conclusions prioritised from low-risk studies and results from moderate- and high-risk studies interpreted cautiously when drawing comparative and translational inferences.

F. Risk Assessment

Table III summarises the overall Risk of Bias scores for 16 studies assessed using PROBAST. This tool evaluates prediction models across four domains (participant, predictor, outcome, and analysis) to identify bias levels. Table III show the Overall ROB using PROBAST. Six studies rated moderate (37%), including both traditional and hybrid ML, models like XGBSE, MT-GRU and Hybrid ML [23]-[24], the models demonstrated high predictive accuracy but lacked in dataset diversity.4 studies rated Low (25%), which included transformer based models like Targeted BEHRT, BEHRT, Hi-BEHRT and Federated BEHRT [10],[24],[20],[25], the models showed a strong data representation and the use of large dataset like MIMIC-II enhanced validation across different data types. Four studies rated Low-Moderate (25%), the models classified were Hybrid, which combine ML and transformers [15]-[16], [13],[26]**.**

For each study that was included, the PROBAST domains (participants, predictors, outcomes, analysis) were scored on a 0-2 scale, which produced an overall bias which ranged from 0(high bias) to 8 (low bias). There were studies scoring >= 6 were classified as low risk, between 4-5 were classified as moderate risk, and <=3 were classified as high risk.

TABLE III
OVERALL RISK OF BIAS

Author	**Overall ROB**	**Percentage**
[10],[24],[20], [25]	Low	25
[23],[27],27], [28],[29],[24]	Moderate	37
[15] , [16], [13],[26]	Low-moderate	25
[30],)[31]	High	13

The studies lacked transparency in Hyperparameter optimisation, which increased the risk of bias. 12% (2 studies) rated high [28], [29], a common factor being that the models used publicly available Kaggle datasets, which lacked longitudinal EHR validation. The PROBAST assessment highlighted improved methodological quality from 2020 to 2025, with the introduction of a hybrid architecture aimed at reducing bias. The 16 studies reveal a need for explainable model frameworks that will aim to reduce bias and facilitate better clinical adoption.

III RESULTS AND DISCUSSIONS

This section represents the characteristics of the included studies. Figure 2 presents the PRISMA diagram of the search and the screening results.

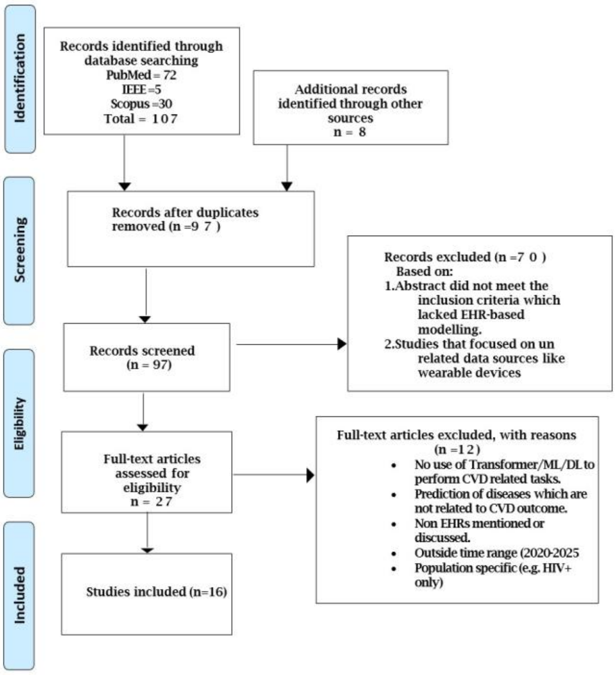
Identification
Screening
Eligibility
Included
Records identified through database searching PubMed = 72 IEEE =5 Scopus =30 Total = 107
Additional records identified through other sources n = 8
Records after duplicates removed (n =9 7)
Records screened (n = 97)
Records excluded (n =7 0) Based on:
1.Abstract did not meet the inclusion criteria which lacked EHR-based modelling.
2.Studies that focused on un related data sources like wearable devices
Full-text articles assessed for eligibility n = 27
Full-text articles excluded, with reasons (n =12)
• No use of Transformer/ML/DL to perform CVD related tasks.
• Prediction of diseases which are not related to CVD outcome.
• Non EHRs mentioned or discussed.
• Outside time range (2020-2025
• Population specific (e.g. HIV+ only)
Studies included (n=16)

Figure 2. PRISMA chart

TABLE IV
SUMMARY OF INCLUDED STUDIES

Author	**Country**	**EHRs source(dataset)**	**ML/Transformer**	**Comparison Models**	**CVD outcomes /Task**	**Opportunities**	**Challenges**	**XAI used**	**Performance metrics**
[23]	United Kingdom	Oxford University Hospital EHR	Multi-Task Gated Recurrent Unit(MT-GRU)/MT-Attention-based(Att)-GRU(RNN)	-QResearch Cardiovascular Risk Algorithm Version 2(QRISK2) , -Logistic Regression(LR) single GRU	-Myocardial Infarction(MI)-Ishaemic stroke	-Longitudinal Electronic Health Record(EHR)	Limited external validation	-Attention weights	-Area Under Curve(AUC)MI- 0.897 -Stroke -0.849
[15]	UK	-Clinical Practice Research Datalink (CPRD)	- Transformer risk model (TRisk)	-QRISK -ML	Cardiovascular disease(CVD) treatment selection	-Improved individualised risk stratification.	Requires validations	Attention visualisation	Concordance-index and AUC improved reporting (-+ 0.1) in papers
[30]	Saudi Arabia	Kaggle	-Ensemble ML -Deep Neural Network(DNN)	-Random Forest(RF) -Extreme Gradient Boosting (XGBoost) -DNN -K-Nearest Neighbours (KNN)	Binary CVD presence	Tabular data enhances a strong benchmark	-Public datasets do not show real clinical data. -Lack of external validation.	RF feature importance	-RF accuracy 88.65%; -AUC (0.92-0.94)
[18]	-UK	-CPRD	Bidirectional Encoder Representation from Transformers for Electronic Health Records(BEHRT) model	-Recurrent Neural Network (RNN) -Long Short-Term Memory(LSTM) -RETAIN -Deep care	CVD onset	The long dependency of models	-Limited external validation -High computational needs	Attention visualisation	Precision gains 8 to 13% over Deep Learning (DL) baselines
[16]	United States of America (USA)	Columbia University Irving Medical Centre-New York Presbyterian Hospital-Observational Medical Outcomes Partnership (CUIMC-NYP OMOP)	Clinical Electronic Health Records (CEHR-BERT)	-BEHRT -MEDBERT -BI-LSTM -XGBoost	Heart Failure(HF)	Temporal reasoning. .	US data only used	Attention component analysis	-AUC 0.80 - 0.84 -Precision Recall (PR)-AUC -0.32
[13]	UK	CPRD	Carefully Optimised and Rigorously Evaluated-Record	-BEHRT -Other transformers	CVD-related risk prediction	Improved calibration	Requires external validation across different institutions.	-No XAI specified, -attention analysis (improved)	High calibration and accuracy vs BEHRT

			(CORE)-BEHRT						
[14]	UK	CPRD	Hierarchical -BEHRT	-BEHRT -CEHR-BERT -RNN -Convolutional Neural Network(CNN)	-HF -Stroke -Chronic Kidney Disease (CKD)	Incorporates multimodal, hence enhancing model efficiency. .	-Architectural complexity -Challenges in deployment	Bootstrap Your Own Latent (BYOL)	-AUROC: Percentage change 1-5% -AUROC PC 1-8 % vs BEHRT
[20]	United Kingdom	Clinical Practice Research Datalink (longitudinal EHR)	Targeted-BEHRT	-Regression -BEHRT -Dragonnet -Targeted Maximum Likelihood Estimation (TMLE)	Estimation of drug effect	Casual inference	Computational intensive	Attention – doubly robust estimation	-Lower Standard Absolute Error (SAE) vs benchmarks; -accurate Relative Risk (RR) estimation
[25]	Israel	Medical Information Mart for Intensive Care-Version III (MIMIC-III)	Federated BEHRT	-Centralised BEHR -Local models	-Next -visit diagnosis prediction (including CVD)	-Training data of the same hospital (near-central performance) -The model can train on several patient datasets**.**	Simulation cost	Attention-based interpretability	-Average precision = 0.63; - within 3% of the central model
[26]	USA	MIMIC-III	BERT-based	-CNN -LSTM -RETAIN	Multi-disease support, including CVD diagnostic	Captures temporal embedding	-Leverages ICU data only -Requires heavy computing	Attention visualisation	-AUC= 0.90 vs CNN =0.84 - LSTM =0.86.
[27]	China	Patient safety and EHR dataset	ML-based predictive model(XGBSE)	-RF -LR -XGBoost	-CVD onset.	-The ML models applicable to clinical settings.	There is a need for external validation due to dataset heterogeneity.	Shapley Additive Explanations (SHAP)	Competitive AUC reported
[32]	China	China Health and Retirement Longitudinal Study (CHARLS) cohort	-KNN -RF -XGB -Light Gradient Boosting (LGB)	LR baseline	-Coronary Artery Disease (CAD) -HF -Angina -Stroke	Transparent pre-processing	Limited recall	SHAP.	-LGB AUC = 0.818 -F1 = 0.509 -Recall = 43.1%
[28]	Turkey	Kaggle	-XGBoost -RF	Support Vector Machine (SVM) -KNN -LR	Binary CVD presence (1 or 0)	Clinical Predictors identified	-Generalisation limits -No longitudinal EHR	SHAP	-XGBoost AUC =0.803 - F1= 0.75
[31]	Turkey	-University of California, Irvine (UCI) heart disease. -Kaggle	Hybrid ML	-LR -RF -XGBoost -SVM -DNN	Binary CVD presence (1 or 0)	Enhanced feature selection pipeline.	-Overfitting risk. -Small non-longitudinal datasets	SHAP	-XGBoost accuracy = 97.4% -AUC =0.98

[29]	USA	Biomarkers integrated cohort	-RF -SVM -Neural Network (NN)	ASCVD(Atherosclerotic Cardiovascular Disease)scores	CVD onset	Biomarkers are integrated with EHR	Limited external validation	Feature importance analyses	AUC = 0.90 to 0.96
[24]	-UK -China	Irregular-timed EHR	Hybrid (ML and Transformer)	-Standard ML -DL	CVD risk prediction	Irregular sampling is addressed.	Difficulties in pre-processing	Temporal embedding and attention	Improved discrimination and calibration

A.Characteristics of Included studies

Table IV shows a summary of 16 studies (study characteristics) from 2020-2025. The included studies employ DL, ML, and transformers as the main predictive models applied to CVD outcomes, leveraging EHRs. Each study is categorised by model type, explainable method used, CVD task, comparative benchmark and performance metrics. Sixteen studies utilised EHRs from Western Healthcare Systems, the United Kingdom (7), China (3), the United States of America (3), and also from Saudi Arabia (1), Israel (1), and Turkey (2). This synthesis highlights the methodological and interpretability techniques, which are guided by the research questions

B. Publication Trends

The studies chosen were from 2020 to 2025; Figure3 shows the trends in the distribution of years.

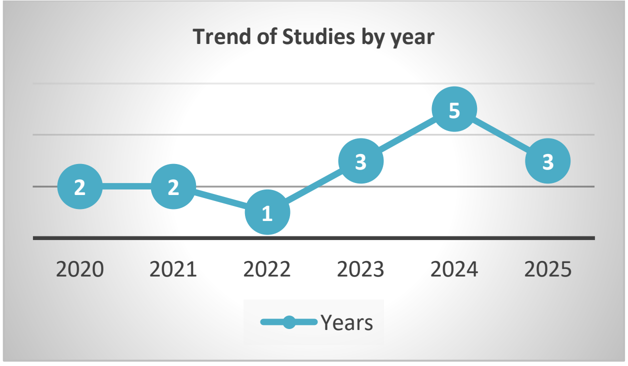

Figure 3. Trend of studies by year

Figure 3 shows the trends in the publication year from 2020 to 2025, the eligible years for the review. In 2020, there were 2 studies; in 2021, there were 2 studies. The least studied was 1 study, 2022. In 2023, there were 3 studies; in 2024, there were five studies, and 3 studies were published in 2025. The studies were published in different countries (Figure 4). The Line chart (Figure 3) represents the rise in research interest in transformer-based and AI-driven prediction models for CVD using EHRs. In 2022, there was a decline, but the number increased from 2023 to 2024, reflecting the growth of Data-driven solutions in healthcare.

C. Study Distribution

The findings were from 6 different countries, as shown in Figure 4 in 2020-2025. The pie chart in Figure 4 shows the distribution of countries in the included studies for review. It shows that 41% of studies (7\16) were from the UK, which dominates the research area, meaning the UK has more active research and better EHR access.

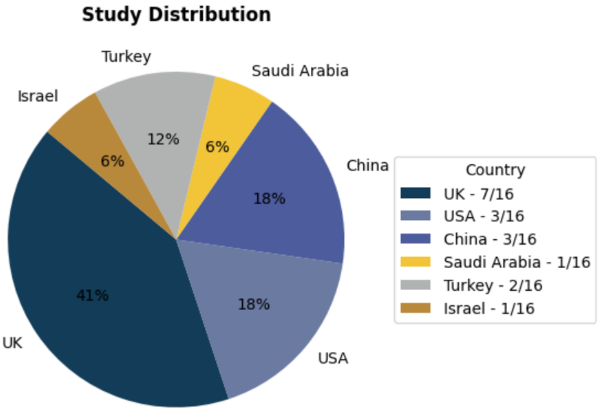

Figure 4. The distribution of studies by country

18%(3) from the USA shows that it is another centre for CVD research on transformers, and China reported 18%(3), which is the second leading, Turkey has 2 studies (12%) highlighting research interest on CVD predictive modelling,6%(1 each) from Saudi Arabia and Israel indicate emerging interest, but less research output compared to the UK or the USA. The distribution indicates the growing interest in Data-driven predictive modelling in healthcare.

D. Performance and Distribution of Model Types

Figure 5 shows the distribution of model types; ML models were most used (9 studies), transformers were applied in 7 studies most used (BEHRT, CORE-BEHRT, Hi-BEHRT, Federated BEHRT and TRisk) [25],[20],[14]. In 2 studies Hybrid approach highlighted the interest in integrating models for temporal reasoning.

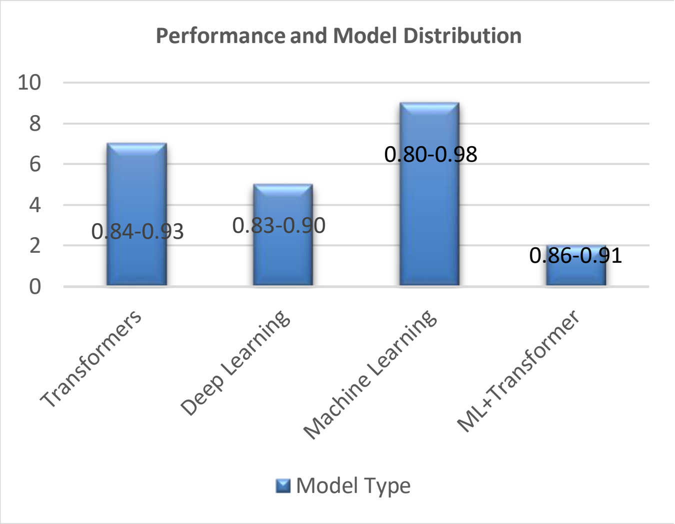

Figure5. Performance and distribution of model type

The reported AUC values ranged from 0.84 to 0.93; however, these variations should be interpreted in light of difference in study design and data characteristics. Higher AUCs were generally observed in studies using curated datasets with limited population diversity, whereas studies relying on real-world EHR data demonstrated a greater performance variability. Differences in longitudinal depth, feature engineering strategies and validation approaches further contributed to the observed performance spread.

Notably, higher AUC values were predominantly observed in studies using curated or single-institution datasets with limited population heterogeneity, whereas studies relying on real-world, longitudinal EHRs exhibited greater performance variability, underscoring the influence of study design and data characteristics on reported discrimination.

E. Model usage and best performance frequency

Figure 6 presents a clustered comparison of model usage across studies and the frequency with which each model achieved the best performance. Transformer-based model demonstrates stronger performance relative to their adoption compared with traditional and deep learning approaches. Among transformers, BEHRT shows the best performance in two studies, being evaluated in three studies.

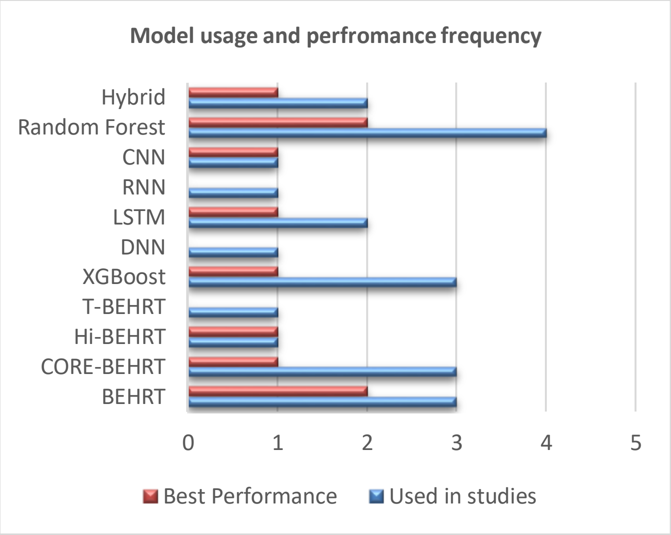

Figure 6. Model Usage

CORE-BEHRT was also used in three studies, but it achieved the best performance in only one, while Hi-BEHRT achieved top performance in its single evaluation. Targeted-BEHRT did not outperform baseline models. These results show BEHRT's effectiveness for longitudinal EHR-based CVD prediction. Traditional machine learning models were frequently used but less consistently dominant. Random forest was a commonly applied ML model that appeared in four studies and performed best in two, which reflected its strength on tabular data but shows limited modelling capacity. XGBoost showed competitive performance but did not exceed the transformer-based models. Deep learning models like LSTM, CNN, RNN and DNN were evaluated in one to two studies each and achieved moderate performance. Lastly, the hybrid ML-Transformer model demonstrated promising results in limited evaluations (1). Figure 6 indicates that transformer-based, particularly BEHRT, achieves superior performance, which supports their growing adoption for CVD prediction from longitudinal EHRs.

F. Cardiovascular Disease Outcomes

The CVD outcomes are presented in Table V.

TABLE V
CARDIOVASCULAR DISEASE OUTCOME

CVD Task	Number of Studies
Stroke	3
CVD outcome	10
Heart Failure	3
Chronic Kidney Disease	1
Drug	1
Coronary Artery Disease	1
Angina	1
Next visit	1

Table V shows that the general Cardiovascular disease (CVD) outcome was mentioned in 10 studies, which is the largest category. This indicates that researchers aimed to build broad CVD risk prediction models using EHRs. Three studies evaluated the use of predictive models in Stroke, which indicates a significant interest in using EHRs for real-time detection of stroke. Similarly, 3 studies evaluated Heart failures using transformer models because HF involves long-term clinical histories where the transformer can model well, whereas specific CVD types like Chronic Kidney Disease (CKD), Drug effects, Coronary Artery Disease (C, Angina, and next visit predictions were done in 1 study, showing that they are not frequently addressed. Studies like [14] evaluated 3 CVD outcomes like HF, Stroke and CKD simultaneously using Hi-BEHRT, which is a transformer-based model which leverages multimodal data(unstructured and structured data) to make predictions. Another study evaluated Heart failure [16] only on CEHR-BERT.

G. Challenges and Opportunities of Transformer Models

TABLE VI
OPPORTUNITIES AND CHALLENGES OF TRANSFORMER MODELS

Transformer Model	Opportunities	Challenges
[18]	• Captures long-term dependencies in EHRs • Interpretability enhanced through attention mechanisms	• High computational cost. • Limited external validation.
[13]	• Improves calibration and stability for clinical adoption. • Enhance interpretability and readiness of BEHRT outputs.	• Limited transparency in Hyperparameter tuning. • Limited external validation.
[14]	• Supports multimodal learning to enhance efficiency in long-term CVD prediction.	• Computational cost. • Challenges in integrating multimodal data for deployment.
[25]	• Privacy is preserved due to local training across institutions. • Decentralised data storage.	• High communication and simulation cost. • Limited external validation across heterogeneous systems.
[20]	• Integrated causal inferences to enhance drug effect estimation	• Computationally intensive for large datasets. • Limited validation across different populations.
[15]	• Employs attention visualisations for clinical interpretability to improve individualised risk stratification and treatment selection.	• Limited validation in diverse and larger cohorts. • Limited interpretability
[33]	• Incorporates age and time embedding for temporal reasoning in prediction.	• Dataset restricted to the US. • Moderate interpretability. • Limited external validation.
[17]	• Enables disease prediction across multiple domains. • Provides a foundation for structured medical data.	• Excludes temporal embeddings (used in BEHRT) • Limited temporal precision.
[19]	• Integrate textual clinical notes and tabular data. • Enables cross-modal learning for improved disease prediction.	• Increased risk of overfitting. • High architectural complexity. • Lacks specific validation for CVD outcomes.

Table VI shows most common challenge (6 counts) is the computational cost or complexity [18],[14], [19], [25],[15], [25], followed by the limited validation, which enhances generalisation of the model for clinical adoption. Models like multimodal [19] had challenges in integrating data (structured and unstructured), overfitting risk was noted, and communication or simulation cost was identified as a challenge in [25], where local training is done on patient EHRs. Lastly,[13]demonstrated a limited transparency in Hyperparameter tuning.

The most noted opportunities are improved predictive performance in 5 models [16]–[18],[14], [19] and enhanced interpretability through attention visualisation techniques.3 models demonstrated their ability to capture long-term or temporal dependencies [16]–[18] that enhanced model predictions. The preservation of privacy and causal inference capabilities are not common as they appear in 1 model each[20], [25]. Multimodal learning[19] is a growing opportunity, but it is still used in a few studies.

H. Explainable AI

All studies (16) included feature importance or Explainable AI or attention techniques to help interpret the outputs, making models more understandable. Figure 7 shows the XAI used in a number of studies.

Attention weights(AW)-1, Temporal embedding-1, Feature importance (1), SHAP (5), Attention visualisation (6) and Bootstrap Your Own Latent(BYOL) -1 studies.

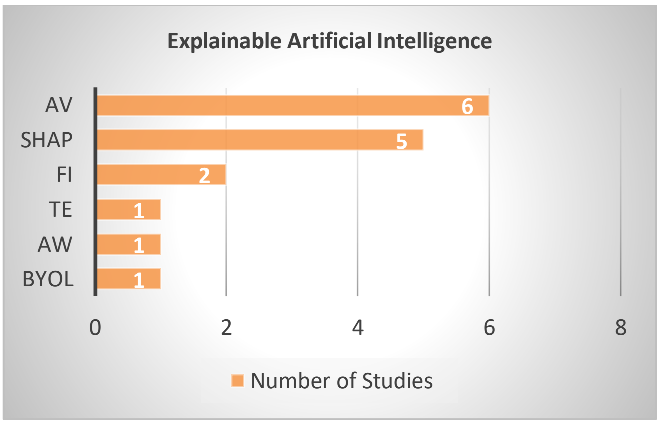

Figure 7. Explainable AI

The chart highlights AV being the most used XAI in the included studies for interpretability. The bar plot (Figure 7) shows that researchers are trying to bridge the gap of the “black box” by demonstrating how the model predicted any CVD task to enhance clinician trust; however, the models are still not deployed in clinical settings.

Across the 16 reviewed studies, three dominant patterns arose, such as: a) the transformer model outperformed DL/ML in capturing the temporal dependencies, even though the model remained inconsistent in external validation and calibration. b) XGBoost was noted to be the highest ML model, which achieved high accuracy on small curated datasets, which highlights inflation in performance. Transformer models were shown to rely on post-hoc attention visualisation, not clinically grounded explanatory mechanisms, so these patterns demonstrated the fragmentation of methodological strategies and lack of evaluation protocols across the reviewed papers.

Overall, the evidence suggests that performance gains alone are insufficient to justify clinical adoption; instead, interpretability, robustness under data shift, and workflow integration emerge as equally critical determinants of real-world utility.

I. Discussion

A critical synthesis of the reviewed papers demonstrates that, despite the discriminative performance of transformer models, they still lack external validations. This section answers the four research questions and is numbered as follows.

1) RQ1 What are the most common algorithms and transformer models that have been applied to Cardiovascular disease prediction leveraging EHRS?

Among the conventional algorithms, XGBoost and Random Forest remain the most used machine learning baselines for benchmarking, which is due to their robustness on tabular EHRs[34], [35]. They are not suitable for longitudinal EHRs due to their limitations in temporal reasoning and lack of contextual embedding [32]

The random forest algorithm is a powerful machine learning algorithm that uses multiple decision trees for classification and regression tasks, and it operates under supervised learning [31],[7], [36]. It is constructed using decision trees, which are generated from random samples of data and the final output is determined by the majority voting among the trees [34]. The RF algorithm was used [30] within an ensemble for Binary CVD detection, utilising the Kaggle dataset with feature importance to analyse the output. Similarly, [31] compared RF with other ML models using the same dataset as [27], highlighting RF as outperforming the ML models.[29] Integrated RF to predict CVD outcomes with biomarker data, and feature importance was used to interpret the predictions.RF is also used in other fields besides healthcare, like in banking, where RF predicts creditworthiness or detects fraud, and in e-commerce, it is utilised to predict customer preferences based on their past behaviour [34].

The strengths of RF are that it performs well with heterogeneous and structured inputs, and the models that use RF are explainable via feature importance or SHAP. RF have a versatility ability for regression and classification, and has reduced overfitting [34]. The challenges with it are that RF involve longer training times due to the number of trees, which slows down real-time predictions [34]. There is no temporal sequencing, which is a limitation for progressive diseases. The author [30] reported RF accuracy to be 88.7% with an AUC ranging from 0.92 to 0.94 for curated datasets.[29] evaluated RF to be the top model with an AUC ranging from 0.90 to 0.96 in cases where biomarkers are integrated.

XGBoost is an optimised gradient boosted tree algorithm that builds trees sequentially, where each tree corrects the errors of the previous ensemble on tabular data. The process is adding trees that model the residuals through tree pruning [35]. The model was a top-performing baseline in studies like [31],[28] for binary CVD classifications with a high AUC (0.98) for curated data. [27], [32] used XGBoost for a longitudinal-oriented task, where they integrated SHAP to visualise the interpretation of feature contributions.

The strengths of XGBoost are effective in distinguishing between classes [35] and showed superiority in discrimination on tabular sets, for example [31], XGBoost achieved the best among other ML baselines with AUC =0.98, and its accuracy was 97% and can be integrated with SHAP or other techniques to provide feature importance.

The challenge is the limited temporal embedding and overfitting on the curated data [35].

Transformers represent a state-of-the-art deep learning architecture that supports NLP tasks such as word/sentence predictions[37]. They have become one of the most influential architectures in artificial intelligence, where computers have achieved great success not only in NLP but also in vision and speech processing [38]. The transformer models adopted by the studies under review are BEHRT[18], which adapts transformer models originating from Natural Language Processing, tokenises diagnoses/medications, and procedures in a patient timeline. BEHRT uses self-attention to allow the model to weigh the importance of past events to predict future outcomes. The original paper [18] introduced BEHRT to predict 301 diseases, CVD included, from CPRD and MIMIC datasets, when compared against LSTM, CNN and other ML baselines.BEHRT improved discrimination and provided understandable explanations. BEHRT was reported [16]to outperform RNN approaches and prior BERT-based models for multi-disease tasks. The strengths reported in BEHRT are gains in precision of 8-13% over DL baselines, enhanced understandability of the model output, and BEHRT is a base for variants, capturing longitudinal trajectories in EHRs [16].BEHRT challenges are the need for large longitudinal datasets, limited deployment in resource-constrained settings, and the need for external validation across institutions.BEHRT reported 0.84-0.93 for the CVDs task, compared to LSTM/CNN.

BEHRT variants (Hi-BEHRT, Federated-BEHRT, Targeted-BEHRT*)*

Hi-BEHRT [14] introduces multimodal inputs, CORE-BEHRT [13] is an evaluated BEHRT that focuses on calibration for clinical use, while Targeted-BEHRT [20] uses causal inference for treatment effect on the individual, and Federated-BEHRT [25] trains the BEHRT model locally on the participating institute's dataset.

- CORE-BEHRT improved calibration [13] to make BEHRT clinically adoptable.
- Hi-BEHRT [14] used multimodal data to predict HF, Stroke and CKD.
- Targeted-BEHRT [20] extended the BEHRT and used causal inference to estimate drug effects within EHRs.
- Federated-BEHRT,[25]trained the BEHRT model on local datasets or devices to maintain data privacy.

Despite promising predictive performance, the clinical implementation of transformer based model for cardiovascular disease prediction remains limited. Most studies have not evaluated integration within routine clinical workflows, real-time inference feasibility or any clinician interactions with the model outputs. Furthermore, the absence of prospective validation and impact assessment restricts the translational readiness of these models, underscoring the need for deployment-oriented evaluation beyond retrospective performance metrics.

2) RQ2 How do the transformer models perform compared to traditional, deep learning and machine learning approaches based on evaluation metrics such as AUC:

The Area under Curve(AUC) is a common metric for assessing the discriminative ability of prediction models [39]. In this study, AUC was commonly used as a performance metric, followed by accuracy [32],[28], F1-Score, Precision, calibration, Concordance-index, Precision-Recall AUC, and Average Precision. Across 16 included papers, transformers (BEHRT, CORE-BEHRT, Hi-BEHRT) achieved AUC (0.84 to 0.93), which shows a higher discrimination and calibration compared to Deep Learning AUC (0.83 to 0.90) and Machine Learning AUC (0.80 to 0.98), which depend on dataset quality[13], [14], [16], [18]. The TRisk model utilised the CPRD dataset, which refined patient stratifications and enhanced treatment allocation, improving by 0.1 over QRISK3 in concordance index[15].

CORE-BEHRT enhanced calibration, which was a bedrock of clinicians' translation [13], while Hi-BEHRT reported improved discrimination (AUCROC = 0.91) among the multimodal datasets [14]. The Targeted-BEHRT[20] reduced the standard error by integrating causal inference for estimating drug effects. Another transformer-based study achieved an AUC of 0.957 for CVD classification by combining with statistical feature filtering [40]. Despite the performance gain in transformers, there is still a generalisation concern; for instance, there is a study that assessed deep learning models under data shifts [41], reported that models like BEHRT capture trajectories, and performance lowered when no stationary data was used [42].In comparison, traditional ML achieved high AUC (0.98) [28] on small curated datasets but had poor generalisation on longitudinal EHRs [30],[31].

While transformer-based architectures offer advantages in modelling long-term temporal dependencies and heterogeneous EHR data, conventional machine learning models such as Random Forest, Gradient Boosting, and Logistic Regression remain competitive for structured tabular datasets. Several reviewed studies reported comparable performance between transformers and non-transformer models when temporal complexity was limited, suggesting that model selection should be guided by data characteristics rather than architectural novelty alone.

3) *RQ3 What explainable artificial intelligence (XAI) techniques have been integrated into Transformer models to improve clinical interpretability?*

Interpretability is vital in clinical modelling [43] to understand how systems make decisions, as it remains a critical issue for clinicians or stakeholders who are involved in the process and affected by the prediction result [44]. Explainable Artificial Intelligence enhances models to be understandable and transparent in diagnostic tasks [45], [46], [47],[48]. Among transformer-based studies, attention weights have been used to show relevant temporal events. Hybrid models integrate SHAP or Feature importance. For example, [49] explored code-level interpretability of the transformer models. Fifteen out of sixteen studies integrated a type of Explainable AI (XAI) tool to enhance model understandability. The frequently used techniques were SHAP and Attention visualisation [25], [29],[28]. The dominance of attention-based methods indicates that interpretability is central to transformer-based modelling, though clinical explainability remains limited [11], [15], [18]. An example is the TRisk model that used attention mechanisms to visualise the risk driving events that are in a patient timeline [13], and BEHRT used bidirectional attention weights for the interpretation of diagnosis [10].SHAP was noted to be used most by ML models to rank the significance of biochemical predictors and their demographic characteristics[28],[28].Hi-BERHT deployed Bootstrap your own Latent(BYOL) for multimodal feature interpretation [14].

However, while attention maps provide some insight, many studies still rely on post-hoc XAI rather than inherently interpretable model structures[3], [8]. Moreover, only a minority of studies translate XAI outputs into actionable clinical insights. A key gap is the translation of model interpretation into clinician-friendly decision support by aligning the clinical interpretability and technical frameworks. Despite the great opportunity demonstrated by the XAI tools, there was limited proof that the interpretability outputs were evaluated in the clinical workflow and from the XAI tools, attention maps dominated in transformer-based studies, offering partial interpretability. The maps may produce misleading explanations when the attention does not correlate with casual drivers. This gap demonstrates a disconnection between clinical explainability and technical interpretability.

4) *RQ4 What limitations and opportunities are presented by transformer models for future empirical investigation?*

Transformer-based models have achieved state-of-the-art results in predicting CVD [3]. However, they face key limitations that hinder their adoption in the clinical sector, including data privacy concerns, as EHRs contain sensitive patient information [6], challenges in integrating multimodal data, even though its modelling has emerged as a powerful approach in clinical research[50], high computational costs, and a lack of external validations[51], [52] across diverse settings [6], [25]. As shown in the summary table IV, studies are concentrated in different countries, often using databases like CPRD, a US-based dataset, which means other populations are underrepresented, and models are not trained on diverse datasets to improve generalisation [29], [9], [11], [12],[53]. Leveraging longitudinal EHRs restricts model adoption or applicability in low-resource settings with inconsistent data [9], [11]. While attention visualisations [9], [10], [12], [15], [25], [26] offer insights into model outputs, interpretability remains a concern due to the lack of transparency required for clinical adoption. Another limitation is the absence of fairness[54] or bias analysis; from the reviewed studies, performance disparities related to socioeconomic groups, sex, and age were not evaluated, despite CVD risk varying across populations[16]. Without fairness assessments, transformer models risk exacerbating existing health inequalities.

Despite the limitations noted, some opportunities for transformer-based models emerged, like multimodal transformers (HI-BEHRT) [14], [55]. The model demonstrated the use of different EHR structures to predict the onset of Heart failure, thereby improving efficacy. Multimodal data analysis is a great emerging opportunity in the healthcare setting as many modalities are considered in the analysis or prediction of a certain risk, which makes models more accurate as all aspects are investigated, such as data from wearable devices, clinical notes and imaging[55]. Another opportunity emerging is federated learning, which stands as a pathway to enable multi-institutional collaboration with centralised storage of data [25]. The ability of transformer-based models to capture long-term dependences in EHRs [18] makes the models outperform ML techniques. The transformer-based model, such as BEHRT, can identify relationships in patient data over time to understand the evolution of the disease under study. The transformers reviewed in studies [15], [16], [25] enhanced interpretability through attention mechanisms where weight is given to each past event, like the medication the patient is subscribed to, which promotes transparency since the clinicians will see the events the model predicted on, and how much weight they each carried. The study [13] demonstrated how transformers improved calibration with their estimation probability more likely to be the same as the real-world events, making it stable for clinical adoption. Privacy of patient EHRs improved by the introduction of federated learning [25], where there was local training across institutions, aimed at decentralising data storage through simulation of the system to enable clinicians to make informed decisions for their patients.

From a clinical implementation perspective, the reviewed evidence suggests that transformer-based models are not yet ready for routine deployment. Most studies remain retrospective, lack prospective evaluation, and do not assess the feasibility of real-time inference or clinician interaction with model outputs. Furthermore, issues related to computational cost, calibration monitoring, and integration into existing clinical workflows remain largely unaddressed, particularly in resource-constrained healthcare settings. These gaps highlight the need for implementation-focused studies that move beyond algorithmic performance toward clinical utility.

J. Limitations of the study

Generalisation across populations remains a significant challenge. Most studies were conducted using data from single institutions or healthcare systems, thereby increasing the risk of population-specific biases and data shift. Variations in demographic composition, clinical coding practices, and healthcare delivery models may substantially reduce model performance when applied to external settings, underscoring the need for multi-centre validation and robustness testing. The concentration of studies in a small number of countries and healthcare systems further amplifies concerns about population generalisability, as models trained on homogeneous cohorts may fail to perform reliably when exposed to demographic, clinical, and institutional shifts.

K. Future work

Future research should focus on four key priorities such as: 1) systematic evaluation of robustness under temporal and distributed data shifts; 2) large-scale external validation across geographically and demographically diverse populations; 3) integration of intrinsic interpretability mechanisms within transformer architectures to support clinical decision making; and 4) simulation-based studies assessing clinical impact and workflow integration.

Together, these priorities constitute a focused research agenda for advancing transformer-based CVD prediction models from experimental success toward clinically deployable decision support systems.

L. Implication of the study

1) Practical implications: Healthcare organisations seeking to implement transformer-based models for the prediction of CVD may consider the model as a state-of-the-art option using longitudinal EHRs, but there should be a plan for computational cost and a strategic way to integrate the models into the clinical workflow. Healthcare providers must account for external validation and calibration monitoring when deploying models in the healthcare sector. Finally, policymakers and funders in healthcare should support federated learning or the standardisation of EHRs to facilitate the adoption of transformer models across all settings.

2) Theoretical implications: The results represent a learning theory by demonstrating how transformer-based models like BEHRT or BEHRT's variants capture temporal dependencies in EHR data, which outperform traditional CNNs and RNNS in modelling longitudinal patient trajectories. The study extends explainable AI by identifying attention mechanisms and SHAP visualisation for model understandability. This reinforces trust in AI systems by demonstrating how the features contributed to the model's output to enhance clinical adoption. Lastly, the review demonstrated the theory of information integration by highlighting multimodal transformers such as Hi-BEHRT that support precision medicine. The present SLR advances theory by demonstrating that sequential attention-based architectures offer a novel approach to modeling the disease progression task. This revealed their weakness in situations marked by irregular sampling, differences across institutions, and missing data.

IV. Conclusion

The SLR highlights the growing potential of transformer-based models for cardiovascular disease prediction from electronic health records. While these models demonstrate strong performance in handling longitudinal and heterogeneous data, current evidence remains largely experimental. Widespread clinical adoption will depend on rigorous external validation, careful consideration of generalisability, and meaningful integration of interpretability and deployment constraints. Consequently, transformer-based approaches should be viewed as promising but not yet ready for deployment in routine cardiovascular risk prediction.

References

[1] B. Chong *et al.*, "Global burden of cardiovascular diseases: projections from 2025 to 2050," *Eur. J. Prev. Cardiol.*, vol. 00, no. 0, pp. 1–15, 2024, doi: 10.1093/eurjpc/zwae281.

[2] World Health Organization, *World health sWORLD HEALTH ORGANIZATION - World health statistics 2024. ISBN 9789240094703. tatistics 2024*. 2024.

[3] E. Antikainen *et al.*, "Transformers for cardiac patient mortality risk prediction from heterogeneous electronic health records," *Sci. Rep.*, vol. 13, no. 1, pp. 1–10, 2023, doi: 10.1038/s41598-023-30657-1.

[4] G. Pallavi, "QFM-BioPred : Quantum Fusion Model for Bioactivity Prediction in Cardiovascular Disease Drug Discovery," no. September, pp. 1–12, 2025, doi: 10.47852/bonviewJCCE52025138.

[5] J. J. Rose *et al.*, "Cardiopulmonary Impact of Electronic Cigarettes and Vaping Products: A Scientific Statement from the American Heart Association," *Circulation*, vol. 148, no. 8, pp. 703–728, 2023, doi: 10.1161/CIR.0000000000001160.

[6] M. Di Cesare *et al.*, "The Heart of the World," *Glob. Heart*, vol. 19, no. 1, 2024, doi: 10.5334/gh.1288.

[7] M. Sibindi, S. Sibanda, J. Luke, and C. Mugwanda, "A Predictive Model for Personalized Healthcare Management for Patients with Chronic Diseases," no. July, 2024, doi: 10.46254/EU07.20240164.

[8] T. Liu, A. J. Krentz, Z. Huo, and V. Ćurčin, "Opportunities and Challenges of Cardiovascular Disease Risk Prediction for Primary Prevention Using Machine Learning and Electronic Health Records: A Systematic Review," *Rev. Cardiovasc. Med.*, vol. 26, no. 4, 2025, doi: 10.31083/RCM37443.

[9] T. Liu, A. Krentz, L. Lu, and V. Curcin, "Machine learning based prediction models for cardiovascular disease risk using electronic health records data: systematic review and meta-analysis," *Eur. Hear. J. - Digit. Heal.*, vol. 6, no. 1, pp. 7–22, 2025, doi: 10.1093/ehjdh/ztae080.

[10] N. W.C Mukura and B. Ndlovu, "Performance Evaluation of Artificial Intelligence in Decision Support System for Heart Disease Risk Prediction," no. Who 2018, pp. 83–93, 2023, doi: 10.46254/ap04.20230043.

[11] Y. Cai *et al.*, "Artificial intelligence in the risk prediction models of cardiovascular disease and development of an independent validation screening tool: a systematic review," *BMC Med.*, vol. 22, no. 1, pp. 1–18, 2024, doi: 10.1186/s12916-024-03273-7.

[12] R. Nadarajah *et al.*, "Prediction models for heart failure in the community: A systematic review and meta-analysis," *Eur. J. Heart Fail.*, vol. 25, no. 10, pp. 1724–1738, 2023, doi: 10.1002/ejhf.2970.

[13] M. Odgaard, K. Klein, S. M. Thysen, E. Jimenez-Solem, M. Sillesen, and M. Nielsen, "CORE-BEHRT: A Carefully Optimized and Rigorously Evaluated BEHRT," *Proc. Mach. Learn. Res.*, vol. 252, pp. 1–33, 2024.

[14] Y. Li *et al.*, "Hi-BEHRT: Hierarchical Transformer-Based Model for Accurate Prediction of Clinical Events Using Multimodal Longitudinal Electronic Health Records," *IEEE J. Biomed. Heal. Informatics*, vol. 27, no. 2, pp. 1106–1117, 2023, doi: 10.1109/JBHI.2022.3224727.

[15] S. Rao *et al.*, "Refined selection of individuals for preventive cardiovascular disease treatment with a transformer-based risk model," *Lancet Digit. Heal.*, vol. 7, no. 6, p. 100873, 2025, doi: 10.1016/j.landig.2025.03.005.

[16] R. Chen and A. Perotte, "Pang21a," pp. 239–251, 2021.

[17] L. Rasmy, Y. Xiang, Z. Xie, C. Tao, and D. Zhi, "Med-BERT: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction," *npj Digit. Med.*, vol. 4, no. 1, 2021, doi: 10.1038/s41746-021-00455-y.

[18] Y. Li *et al.*, "BEHRT: Transformer for Electronic Health Records," *Sci. Rep.*, vol. 10, no. 1, pp. 1–12, 2020, doi: 10.1038/s41598-020-62922-y.

[19] M. Mbaye and M. Danziger, "Multimodal BEHRT : Transformers for Multimodal Electronic Health Records to predict breast cancer prognosis," 2024.

[20] S. Rao *et al.*, "Targeted-BEHRT: Deep Learning for Observational Causal Inference on Longitudinal Electronic Health Records," *IEEE Trans. Neural Networks Learn. Syst.*, vol. 35, no. 4, pp. 5027–5038, 2024, doi: 10.1109/TNNLS.2022.3183864.

[21] M. J. Page *et al.*, "The PRISMA 2020 statement: an updated guideline for reporting systematic reviews," 2021, doi: 10.1136/bmj.n71.

[22] M. Milne-Ives, E. Selby, B. Inkster, C. Lam, and E. Meinert, "Artificial intelligence and machine learning in mobile apps for mental health: A scoping review," *PLOS Digit. Heal.*, vol. 1, no. 8, pp. 1–13, 2022, doi: 10.1371/journal.pdig.0000079.

[23] F. Andreotti *et al.*, "Prediction of the onset of cardiovascular diseases from electronic health records using multi-task gated recurrent units," 2020.

[24] C. Li *et al.*, "Improving cardiovascular risk prediction through machine learning modelling of irregularly repeated electronic health records," *Eur. Hear. J. - Digit. Heal.*, vol. 5, no. 1, pp. 30–40, 2024, doi: 10.1093/ehjdh/ztad058.

[25] O. Ben Shoham and N. Rappoport, "Federated learning of medical concepts embedding using BEHRT," *JAMIA Open*, vol. 7, no. 4, pp. 1–10, 2024, doi: 10.1093/jamiaopen/ooae110.

[26] R. Tang *et al.*, "Embedding Electronic Health Records to Learn BERT-based Models for Diagnostic Decision Support," *Proc. - 2021 IEEE 9th Int. Conf. Healthc. Informatics, ISCHI 2021*, pp. 311–319, 2021, doi: 10.1109/ICHI52183.2021.00055.

[27] K. Lee, S. W. Oh, S. H. Kim, T. Ko, and I. Y. Choi, "Machine Learning-Based Predictive Models for Early Detection of Cardiovascular Diseases: A Study Utilizing Patient Samples from a Tertiary Health Promotion Center in Korea," *Stud. Health Technol. Inform.*, vol. 316, no. Ml, pp. 710–711, 2024, doi: 10.3233/SHTI240512.

[28] K. K. Kırboğa and E. U. Küçüksille, "Identifying Cardiovascular Disease Risk Factors in Adults with Explainable Artificial Intelligence," *Anatol. J. Cardiol.*, vol. 27, no. 11, pp. 657–663, 2023, doi: 10.14744/AnatolJCardiol.2023.3214.

[29] Sadia Latif, Sami Ullah, Aafia Latif, Ghazanfar Ali, Muhammad Hassnain Azhar, and Salman Ali, "Predictive Modeling of Cardiovascular Disease Using Machine Learning Approach," *Kashf J. Multidiscip. Res.*, vol. 2, no. 02, pp. 207–232, 2025, doi: 10.71146/kjmr288.

[30] A. Alqahtani, S. Alsubai, M. Sha, L. Vilcekova, and T. Javed, "Cardiovascular Disease Detection using Ensemble Learning," *Comput. Intell. Neurosci.*, vol. 2022, 2022, doi: 10.1155/2022/5267498.

[31] A. H. Elmi, A. Abdullahi, and M. A. Barre, "A machine learning approach to cardiovascular disease prediction with advanced feature selection," *Indones. J. Electr. Eng. Comput. Sci.*, vol. 33, no. 2, pp. 1030–1041, 2024, doi: 10.11591/ijeecs.v33.i2.pp1030-1041.

[32] Q. Huang *et al.*, "Characterisation of cardiovascular disease (CVD) incidence and machine learning risk prediction in middle-aged and elderly populations: data from the China health and retirement longitudinal study (CHARLS)," *BMC Public Health*, vol. 25, no. 1, 2025, doi: 10.1186/s12889-025-21609-7.

[33] R. Chen and A. Perotte, "CEHR-BERT: Incorporating temporal information from structured EHR data to improve prediction tasks," *Mach. Learn. Heal.*, vol. 158, pp. 239–251, 2021.

[34] H. A. Salman, A. Kalakech, and A. Steiti, "Random Forest Algorithm Overview," vol. 2024, pp. 69–79, 2024.

[35] K. Budholiya, S. K. Shrivastava, and V. Sharma, "An optimized XGBoost based diagnostic system for effective prediction of heart disease," *J. King Saud Univ. - Comput. Inf. Sci.*, vol. 34, no. 7, pp. 4514–4523, 2022, doi: 10.1016/j.jksuci.2020.10.013.

[36] S. Hadebe, B. Ndlovu, and K. Maguraushe, "Managing Diabetes Using Machine Learning and Digital Twins," *Indones. J. Innov. Appl. Sci.*, vol. 5, no. 2, pp. 145–162, 2025, doi: 10.47540/ijias.v5i2.1981.

[37] C. A. Siebra, M. Kurpicz-Briki, and K. Wac, *Transformers in health: a systematic review on architectures for longitudinal data analysis*, vol. 57, no. 2. Springer Netherlands, 2024.

[38] T. Lin, Y. Wang, X. Liu, and X. Qiu, "A survey of transformers," *AI Open*, vol. 3, pp. 111–132, Jan. 2022, doi: 10.1016/J.AIOPEN.2022.10.001.

[39] A. C. J. W. Janssens and F. K. Martens, "Reflection on modern methods: Revisiting the area under the ROC Curve," *Int. J. Epidemiol.*, vol. 49, no. 4, pp. 1397–1403, 2020, doi: 10.1093/ije/dyz274.

[40] P. Dubey and P. Dubey, "Advancing CVD Risk Prediction with Transformer Architectures and Statistical Risk Factor Filtering," 2025.

[41] A. Subbaswamy and S. Saria, "From development to deployment: dataset shift, causality, and shift-stable models in health AI," *Biostatistics*, vol. 21, no. 2, pp. 345–352, 2020, doi: 10.1093/biostatistics/kxz041.

[42] Y. Li *et al.*, "Validation of risk prediction models applied to longitudinal electronic health record data for the prediction of major cardiovascular events in the presence of data shifts," *Eur. Hear. J. - Digit. Heal.*, vol. 3, no. 4, pp. 535–547, 2022, doi: 10.1093/ehjdh/ztac061.

[43] M. Frasca, D. La Torre, G. Pravettoni, and I. Cutica, "Explainable and interpretable artificial intelligence in medicine: a systematic bibliometric review," *Discov. Artif. Intell.*, vol. 4, no. 1, 2024, doi: 10.1007/s44163-024-00114-7.

[44] H. N. Cho *et al.*, "Explainable predictions of a machine learning model to forecast the postoperative length of stay for severe patients: machine learning model development and evaluation," *BMC Med. Inform. Decis. Mak.*, vol. 24, no. 1, pp. 1–16, 2024, doi: 10.1186/s12911-024-02755-1.

[45] Z. Revesai, K. Maguraushe, B. Ndlovu, and S. Dube, *Predictive Modeling and Risk Assessment of Monkeypox Transmission Using Bayesian Networks : An Interpretable Machine Learning Approach*. Springer Nature Singapore, 2026.

[46] O. Mabikwa, B. Ndlovu, and K. Maguraushe, "A Comparative Analysis of Machine Learning Techniques and Explainable AI on Voice Biomarkers for Effective Parkinson ' s Disease Prediction,"

vol. 7, no. 3, pp. 2196–2228, 2025, doi: 10.51519/journalisi.v7i3.1172.

[47] B. Ndlovu, K. Maguraushe, and O. Mabikwa, “Machine Learning and Explainable AI for Parkinson’s Disease Prediction: A Systematic Review,” *Indones. J. Comput. Sci.*, vol. 14, no. 2, 2025, doi: https://doi.org/10.33022/ijcs.v14i2.4837.

[48] T. Ngwazi and B. Ndlovu, “Early Detection of Diabetic Retinopathy Through Explainable AI Models : A Systematic Review,” *Int. J. Informatics Dev.*, vol. 14, no. 2, pp. 616–628, 2025, doi: 10.14421/ijid.2025.5200.

[49] E. H. Houssein, R. E. Mohamed, G. Hu, and A. A. Ali, “Adapting transformer - based language models for heart disease detection and risk factors extraction,” *J. Big Data*, 2024, doi: 10.1186/s40537-024-00903-y.

[50] M. Farhadizadeh *et al.*, “A systematic review of challenges and proposed solutions in modeling multimodal data,” pp. 1–46, 2025.

[51] K. I. E. Snell *et al.*, “External validation of prognostic models predicting pre-eclampsia : individual participant data meta-analysis,” pp. 1–18, 2020.

[52] K. I. E. Snell *et al.*, “External validation of clinical prediction models: simulation-based sample size calculations were more reliable than rules-of-thumb,” *J. Clin. Epidemiol.*, vol. 135, pp. 79–89, 2021, doi: 10.1016/j.jclinepi.2021.02.011.

[53] Ghassemi *et al.*, “A Review of Challenges and Opportunities in Machine Learning for Health.,” *AMIA Jt. Summits Transl. Sci. proceedings. AMIA Jt. Summits Transl. Sci.*, vol. 2020, pp. 191–200, 2020.

[54] A. Foryciarz, S. R. Pfohl, B. Patel, and N. Shah, “Evaluating algorithmic fairness in the presence of clinical guidelines: The case of atherosclerotic cardiovascular disease risk estimation,” *BMJ Heal. Care Informatics*, vol. 29, no. 1, 2022, doi: 10.1136/bmjhci-2021-100460.

[55] W. Lyu *et al.*, “A Multimodal Transformer: Fusing Clinical Notes with Structured EHR Data for Interpretable In-Hospital Mortality Prediction.,” *AMIA ... Annu. Symp. proceedings. AMIA Symp.*, vol. 2022, no. Mlm, pp. 719–728, 2022.