Journal of Applied Informatics and Computing (JAIC)
Vol.10, No.1, February 2026, pp. 90~105
e-ISSN: 2548-6861

Transformer-Based Models for Electronic Health Records and Omics in Healthcare: A Systematic Literature Review

Joshua Machemedze [1]*, Belinda Ndlovu [2]**
[1,2] Informatics Department, National University of Science and Technology, Bulawayo, Zimbabwe
joshuacharlesmachemedze@gmail.com [1], belinda.ndlovu@nust.ac.zw [2]

Article Info

Article history:

Received 2025-11-30
Revised 2026-01-10
Accepted 2026-01-17

Keyword:

Artificial Intelligence,
Machine Learning,
Transformer,
Electronic Health Records (EHRs),
Electronic Medical Records (EMRs),
Omics.

ABSTRACT

Electronic Health Records (EHRs) have become central to modern healthcare. The emergence of transformer-based models has profoundly influenced how EHRs are used for modelling complex, longitudinal data. Integration with omics technologies improves the precision of disease identification and risk assessment during modelling. While several reviews have examined transformers in healthcare broadly, a systematic synthesis focused on their architectural design, empirical performance and integration of EHRs with omics data remains limited. This study presents a systematic literature review of transformer-based models applied to electronic health records (EHRs) and omics data, and of their integration into healthcare. Following PRISMA guidelines, peer-reviewed studies were retrieved from IEEE Xplore, ACM Digital Library, PubMed, and ScienceDirect, resulting in 14 eligible empirical studies published between 2020 and 2025. The review analyses transformer architectures, submodules, application domains, comparative performance, interpretability mechanisms, and limitations. Findings indicate that architectural design drives task-specific advantages in disease prediction, phenotyping, medication recommendation, and omics analysis. The integration of self-attention with deep learning, temporal modelling, and a pre-trained biomedical transformer improves performance. However, most studies remain centred on EHR, with limited empirical integration of omics data. Persistent challenges include limited generalisability, high computational cost, data quality issues, and insufficient interpretability for clinical deployment. The primary contribution of this review lies in synthesising architectural trends and methodological gaps. By consolidating current evidence, the study provides clear directions for the development of explainable, generalisable, and multimodal transformer-based systems in precision healthcare.

I. INTRODUCTION

Electronic Health Records (EHRs), introduced in healthcare in 2009, are patient clinical records outside of a particular healthcare provider [1], [2], [3], [4]. This differentiates EHRs from Electronic Medical Records (EMRs), which only contain patient information from a healthcare provider [5], [6]. EHRs enhanced the understanding of human biology and they bridge the gap to precision medicine [7].

To complement EHRs, omics emerged [8]. Omics studies genomics, proteomics, transcriptomics, epigenomics and metabolomics [9], [10], [11]. The technology-enhanced understanding of molecular functions, pathways and interactions[10], [12]. Using Machine Learning, researchers have advanced the level at which this data is analysed to provide insights[13], [14].

Machine Learning (ML), a branch of Artificial Intelligence(AI), learns patterns and generates insights from analysing data [15], [16] . Its subfield, Deep Learning, mimics human cognition using artificial neural networks

[17]. It extends to transformer models, which use self-attention as a mechanism to replace recurrence, allowing global dependency modelling and parallelised training [18]. BERT, BEHRT and Large Language models are transformer models that allow integration of sequential and multimodal data[17], [19]. With transformers, models learn linguistic patterns that traditional AI cannot. Their self-attention attribute enables analysis of sequence data in parallel, considering relationships between attributes irrespective of position in the sequence.

Despite the growing body of reviews on transformer models in healthcare, important gaps remain. Existing systematic reviews primarily focus on natural language processing tasks, longitudinal EHR modelling, or general multimodal applications, often treating omics data peripherally or conceptually [20], [21], [23], [24]. Moreover, prior reviews seldom provide a detailed architectural analysis linking transformer design choices such as attention mechanisms, submodules, and hybrid configurations to empirical performance, interpretability, and clinical applicability. This review addresses these gaps by systematically synthesising empirical evidence on transformer architectures applied to EHRs and omics data, critically comparing their competitive advantages and limitations, and explicitly examining challenges related to generalisability, interpretability, and multimodal integration in healthcare contexts.

Research Questions

1. How are transformer-based models designed, including their core components and submodules?
2. How are recent transformers being applied in different omics and EHRs analytical domains?
3. What are the competitive advantages of each identified transformer model?
4. How can limitations of transformer architectures be addressed in a healthcare context?
5. How do transformers perform compared to DL techniques and traditional modelling techniques?

II. METHODS

The study followed the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines for transparency and completeness [25]. The guidelines included identification, screening and eligibility criteria for the Literature Review.

B. Search Strategy

A comprehensive search was conducted on 13 October 2025 across the IEEE, PubMed, AMC Digital Library, and ScienceDirect databases. A string of keywords and their synonyms was used to filter for papers for this study. ("Machine Learning" OR "Transformer" OR "Deep Learning") AND ("Electronic Health Records" OR "EHR") AND ("Omics" OR "Genomics" OR "Proteomics") was used for Science Direct and "machine learning" OR "ML" OR "deep learning" OR "transformer models") AND ("electronic health records" OR "EHR" OR "electronic medical records" OR "EMR") AND ("omics" OR "multi-omics" OR "genomics" OR "proteomics" OR "transcriptomics" OR "metabolomics") for the rest. The search strategy was designed to maximise coverage while keeping its relevance.

C. Inclusion and Exclusion Criteria

Studies were screened using an inclusion and exclusion criterion, ensuring methodological rigour and relevance. Only papers from 2020 to date are included. This was to ensure the use of the latest research with major transformer models adopted in healthcare research. Peer-reviewed journal articles were selected from recognised databases to avoid grey literature and preprints. The studies were further screened based on title, abstracts and full text evaluation. The inclusion and exclusion criteria are shown in Table I.

TABLE I
INCLUSION AND EXCLUSION CRITERIA

Criteria	Inclusion	Exclusion
Time Frame	2020 to 2025 papers	All papers from 2019 and below
Language	English	All papers not in the English language
Type of Paper	Journal Articles and Conference Papers	Books, book chapters, Systematic Literature Reviews, Grey Literature
Research area	Studies focused on Transformer models or self-attention in precision medicine, omics and EHRs, healthcare	Studies that do not focus on the use of transformers in healthcare
Study Type	Papers with empirical applications of transformers on EHRs or Omics	Papers with theoretical application of transformers in Omics studies, EHRs, or both.

D. Screening

The initial search pulled 76 records from IEEE, 232 from ACM Digital Library, 202 from PubMed and 1008 from Science Direct. The search retrieved 1578 records, which were imported into Mendeley. After the initial search, 2 duplicate papers and 1 Spanish paper were removed. Journals were screened, removing 306 Systematic Literature reviews. An additional n = 846 papers were screened out based on their title. After title screening, the

papers were evaluated based on their abstracts. 234 papers were excluded, leaving 29 documents for the eligibility check. These 29 journals underwent a full-text review. The inclusion of 14 studies reflects the limited empirical research applying transformer architectures to EHR and omics data. It shows the emerging research jointly considering both modalities.

E. Eligibility Criteria

The review examined empirical studies published between 2020 and 2025. The studies under review were peer-reviewed and focused on the use of transformer models in healthcare, using either EHRs, omics, or both modalities. Studies that used AI models other than transformers, or that employed data types other than images, were not eligible for review. Non-peer-reviewed papers were also excluded from the study.

F. Included

During the full-text review, 15 additional papers were excluded because they focused on images or were conceptual or non-transformer-based. Only 14 studies met the study's inclusion criteria. They were empirical studies of transformer models developed for both EHRs and omics, or for a single modality.

G. Quality Assessment

A structured quality assessment was carried out to evaluate the strength, transparency, and scientific rigour of the included studies. The goal was to ensure that only high-quality, evidence-driven papers contributed to the final analysis of transformer-based models applied to EHRs and omics data. The Weighted Technical Methodology (WTM) framework, adopted from multi-criteria decision analysis [26], [27], [28].

This approach measured each study's methodological soundness, data reliability, reproducibility, and interpretability. Five key criteria (C1-C5) guided this assessment, with each scored on a scale of 0-2 and weighted according to its importance in determining the overall quality: model description and reproducibility problem, framing and study design, interpretability and applicability, data quality and appropriateness and clinical validation strategy. The studies were given a weighted score (0–100). Each study was marked as high when equal to or greater than 80, moderate when between 60 and 79, or low quality when less than 60. Risk-of-quality stratification was used to contextualise the synthesis. Conclusions prioritise evidence from high-scoring studies (≥80) and interpret findings from moderate-quality studies (60–79) cautiously, while low-quality evidence (<60) is used only to indicate emerging directions.

III. Results And Discussion

Screening and eligibility results are presented in Figure 1. A 10-column table was constructed to explore research questions and summarise studies eligible for review.

Table II presents a comprehensive review of 10 studies that met the inclusion criteria for this study. It summarises studies by region of origin, dataset type, and primary application areas in the health domain. The table also presents the type of transformer used and its submodules. It also explores the advantages and limitations of the transformer's design, compares it with other models, and outlines the evaluation metrics used in the study.

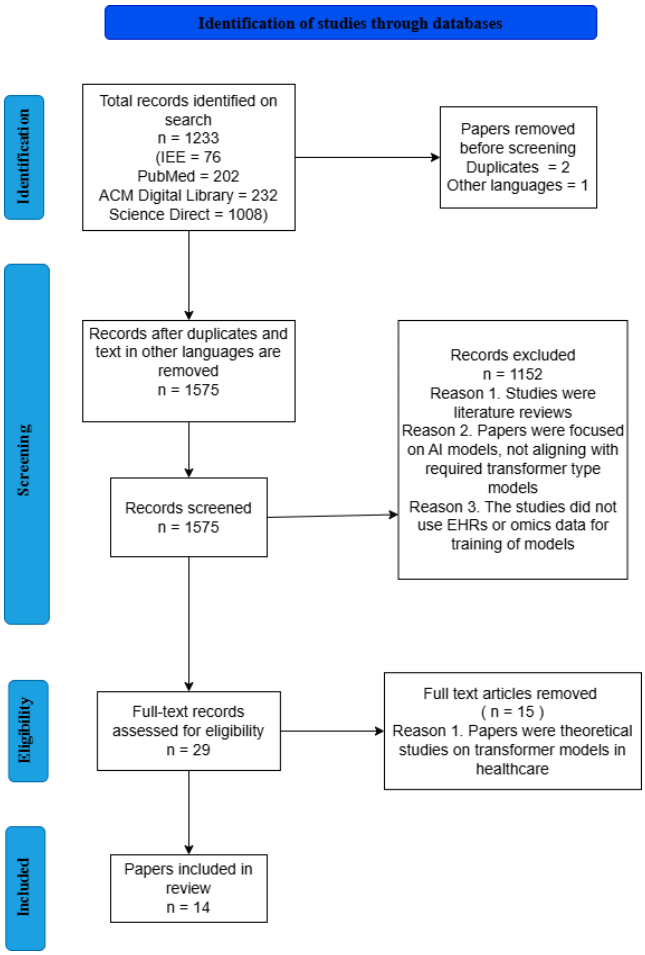

Figure 1. PRISMA screening results

TABLE II

CHARACTERISTICS OF INCLUDED STUDIES

Study	Country	Dataset Type	Primary Application	Transformer Architecture (core + submodules)	Competitive Advantages	Limitations	Baselines Compared
[2]	USA	• EHR	• Disease onset prediction • Patient stratification • Phenotype discovery	• Transformer encoder • Sentence-BERT architecture • Longitudinal embeddings	• Enables unsupervised embedding of EHR sequences • Discovers new comorbidity patterns • Improves forecasting	• Data heterogeneity • Missing modalities • Limited to diagnosis and procedure codes	• BEHRT • Autoencoder
[29]	China	• Event-sequence data (applied to EHR,)	• Sequence modelling • Event prediction	• Universal Transformer • Hawkes Process • ACT mechanism • CNN-enhanced feed-forward layers	• Combines self-attention and recurrence for long-term temporal dependencies	• Overcomes RNN vanishing gradient • Improves event modelling for asynchronous data	• RMTPP • NHP • THP
[30]	Canada	• Continous timeseries (bio-signals) • EHRs	• Extract contextualised representation from timeseries • Learn temporal classification	• TimelyGPT • xPos embedding • Recurrent attention • Temporal convolution modules	• Recurrent retention for forecasting irregularly-sampled time series • Forecasts long sequences of time series	• Permutation invariance of self-attention loses temporal information • Unidirectional architecture • Limited analysis of EHRs	• AutoFormer • TS2Vec
[31]	USA	• Biomedical literature (gene-disease mentions)	• Relation extraction	• BioBERT fine-tuning for relation extraction • KG construction	• Improved precision	• No EHR needed • Focuses on genomics knowledge curation	• TF-IDF clustering • Statistical ML baselines
[32]	China	• Genomics • MRI • Proteomics	• Predict stroke recurrence	• LNet Transformer layer • Dynamic weighting fusion	• Improves cross-modality fusion • Improves performance	• EHR limited • Focus on multi-omics signals	• CNN • SVM • RF baselines
[1]	China	• EHRs • (MIMIC-III and MIMIC-IV	• Medication recommendation • Minimisation of DDI	• Parallel CNN • Transformer encoder (CAT) • GAT over HER • DDI graphs • Joint BCE+DDI loss	• Captures local (visit-level) and long-term (sequential) patterns • Explicit DDI safety	• EHR-only scope • Scalability	• DMNC • RETAIN • LEAP • GAMENet, • MICRON • COGNet • Trans-GAHNet
[33]	United Kingdom	• Longitudinal EHR	• 10-year CVD risk prediction	• BEHRT-derived encoder • Age and encounter embedding • Survival layer	• Strong subgroup generalisation	• Requires full longitudinal EHRs • Limited interpretability	• QRISK3 • DeepSurv • Cox models

Study	Country	Dataset Type	Primary Application	Transformer Architecture (core + submodules)	Competitive Advantage	Limitations	Baselines Compared
[34]	UK	• Linked longitudinal EHRs	• Patient subtyping and prognosis	• Transformer encoder with contrastive learning • Clustering downstream	• Learns disease trajectories • Robust subtypes with prognostic separation	• Missing data • Generalisability issues	• TF-IDF clustering • Statistical ML baselines
[35]	USA	• EHRs	• Modelling phenotypic concepts from diagnosis codes	• RarePT • Masked Language Modelling	• Recapitalate rare diagnosis • Weighting and masked modelling for generalization	• Reliance on ICD-10(noisy, inconsistent) • Uses phecodes which are for common diseases	• Rule based models •
[36]	US + Israel	• Free-text triage notes • Tabular EHR	• Admission risk prediction from triage notes	• Bio-Clinical-BERT fine-tuning • Classification head	• Improves AUC over classic models • Pragmatic compute discussion	• Generalisability	• BOW-LR-TFIDF • W2V-BiLSTM • XGBoost
[37]	UK	• EHRs • Antibiotic administration time-series	• Predict antimicrobial resistance	• 1-D Transformer • Integrated Gradients explanation	• Faster than genomics • Interpretable signatures • Multi-label support	• Handles missing labels • Real-time EHR usage	• Traditional ML baselines
[38]	China	• Multi-omics	• SLE and Lupus Nephritis diagnosis	• Single-head Transformer attention • MLP encoder • Tensor-based bimodal fusion	• Interpretable biomarkers • Generalisation	• Small cohort study • High computational complexity	• SVM • LR • KNN • MOGONET • TEMINET
[39]	USA	• EHRs • Clinical Notes	• Chronic cough prediction	• ClinicalBERT encoder • Custom interpretability attention layer	• Interpretable • Multimodal EHR handling	• No time-gap modelling • Relies on symptom extraction	• LR • SVM • kNN • BiLSTM-Attention • BERT
[40]	Pakistan and Saudi Arabia	• Gene expression • Drug molecular descriptors	• Glioblastoma drug resistance prediction	• Hybrid CNN • BiLSTM • Transformer pathway	• Captures spatial, sequential, contextual signals	• High training complexity • Limited interpretability • Overfitting risk	• CNN • LSTM • Transformers • Decision Trees

A. *Study Origin*

Figure 2 shows the adoption of research on transformer models across continents.

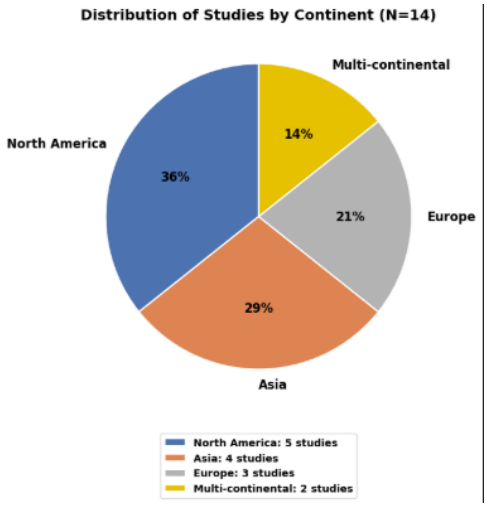

Figure 2. Study Origin

Figure 2 highlights North America as the leading continent. 36% of the new models are being developed on the continent. Asia comprises 29%, 4 out of 14, of the empirical studies in the last 5 years. Europe contributes 21% to the research, and the last 2 new models were developed as a multi-continental collaboration. This indicates the importance and pursuit of generalizable models within the field by using diverse datasets from different continents. Africa reported zero publications or contributions to the research, showing limited adoption of the transformer trend towards personalised medical care. These findings suggest a strong concentration of research capacity in high-resource regions.

B. *Dataset Type*

Figure 3 illustrates the types of datasets used with models in each of the studies included in this study.

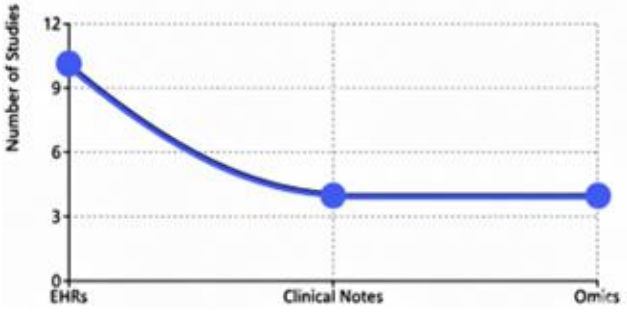

Figure 3. Dataset Type

Among the 14 reviewed studies, 10 used EHRs in model training. This shows that EHRs are the most used and accessible patient data within the research. Four studies used text-based notes, displaying NLP tasks and their importance in learning patterns and deriving insights from EHRs.

However, omics data were used in four studies, which highlights less research on the analysis of bio-mechanisms and functions.

The dominance of EHR-centric datasets highlights a structural limitation in current transformer-based healthcare research. While EHRs provide accessible longitudinal data, their use in isolation constrains biological interpretability and limits the ability of models to capture molecular mechanisms underlying disease. Studies incorporating omics data demonstrated improved diagnostic specificity and biomarker relevance; however, these benefits were offset by increased computational complexity and smaller cohort sizes. This trade-off suggests that current transformer architectures are not yet optimally designed for scalable omics integration, reinforcing the need for architectural innovations that balance performance with feasibility.

Empirical integration of omics remains sparse and methodologically challenging. The reviewed studies report difficulties in synchronising longitudinal EHR events with high-dimensional omics profiles, managing modality heterogeneity (sparse codes vs. dense molecular features), and controlling overfitting under extreme dimensionality. These constraints explain the predominance of EHR-only pipelines and underscore the need for representation learning and alignment strategies tailored to multimodal fusion.

C. *Primary Application*

Figure 4 presents the primary applications of the transformers in this study.

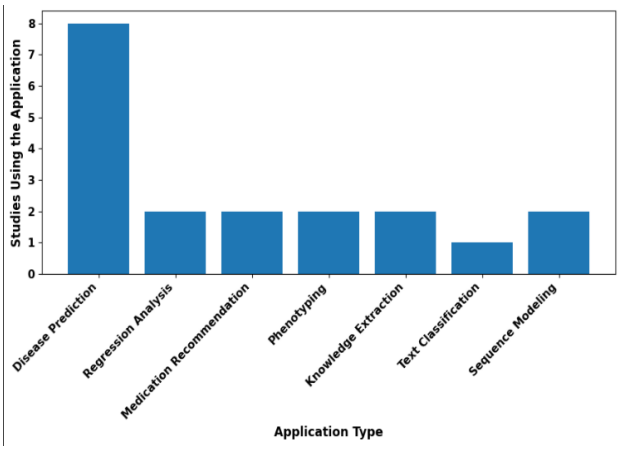

Figure 4. Primary Applications

Eight studies focus on disease prediction and risk forecasting. They prove that transformers are most used for prediction modelling. Two of the studies emphasise patient subtyping or progression analysis. This shows the growing use of embeddings to uncover disease direction and patient groups. Two other papers target medication

recommendation and therapy optimisation, reflecting interest in safer and more interpretable clinical decision support. Two studies apply transformers for phenotyping and knowledge extraction, bridging biomedical NLP with clinical informatics, while two explore sequence modelling, highlighting temporal interpretability.

Overall, prediction-driven research dominates the field, with interpretability and multi-dimensional integration emerging as promising yet still underrepresented directions.

D. Transformer Architecture and Submodules

Figure 5 shows the architectural designs used on the 14 transformers in the included studies.

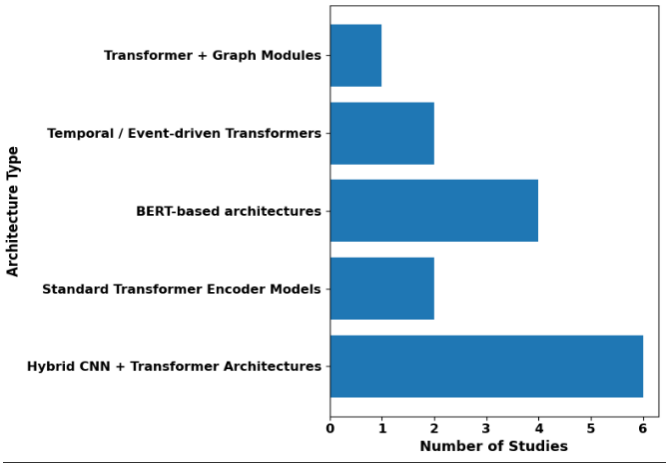

Figure 5. Transformers Architecture and Submodules

Six of the models employ a combination of both CNN and transformer architectures. This highlights a trend in adopting DL and transformer methods. CNN have a single layer, making it good at local dependencies for diverse medical events, and transformers are better at global dependencies and context learning.

Four studies used BERT-based architectures, underscoring the dominance of language models in healthcare for tasks like medical text understanding and clinical reasoning. This indicates the importance of pretrained models in acquiring a holistic understanding of data.

Two other models incorporated standard transformer encoders, and two more employed temporal driven transformers. This highlights how attention layers are favoured in extracting features from input data and how the prediction of health events using longitudinal and time series data is rising in research. One model is based on a transformer and knowledge graphs.

E. Limitations

Limitations of the included studies are summarised in Figure 6.

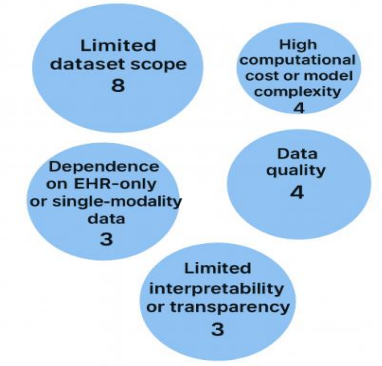

Figure 6. Limitations

Across the 14 studies, eight mention limited generalizability as their main challenge. Most rely on EHR data from one hospital or a small region. Three of the studies highlight the dependence on EHR-only or single-modality data, meaning they do not yet combine genetic and other complementary sources. Four studies highlight the high computational cost of large transformer models, especially when using BERT architectures. Four studies underline noisy or incomplete data as a major drawback, while three mention interpretability as a major limitation. Overall, most papers show impressive results but still face barriers to scaling and applying their models widely in real clinical settings.

F. Competitive Advantage

Table 3 shows the competitive advantages of every architectural design included in this study.

TABLE III.
COMPETITIVE ADVANTAGE

Advantage	**Studies Highlighting the Advantage**
Captures long-term and sequential patterns in EHR data	5
Improves disease prediction and patient representation	2
Enhances interpretability and clinical insight	4
Multi-modal Learning and Generalisation	4
Improves text understanding and extraction of clinical meaning	2
Provides robustness to noisy or missing data	2

Among the 14 studies reviewed, transformers' ability to capture long-term and sequential patient patterns was addressed in five studies. This helps researchers understand disease progression over time. Two studies highlighted their strength in disease prediction and patient representation. This shows how transformer embeddings provide more detailed and context-aware insights. Four studies focused on interpretability, where attention layers made predictions easier to explain using XAI modules in their architectures. Four studies also showed that transformers handle multimodal data well and demonstrated stronger performance in medical text understanding. Two studies also mentioned robustness to missing or noisy data, and two noted faster and more efficient training.

Although transformer models demonstrate clear advantages in sequential modelling and contextual representation, these benefits are not uniform across applications. Performance gains were most pronounced in tasks involving irregular temporal patterns and multimodal inputs, while improvements over deep learning baselines were marginal for simpler tabular EHRs tasks. This variability indicates that architectural suitability, rather than the transformer paradigm itself, largely determines performance outcomes.

G. Compared Baseline Models

Across the studies, researchers compared transformer models to a wide range of baseline approaches. Figure 7 illustrates the compared models.

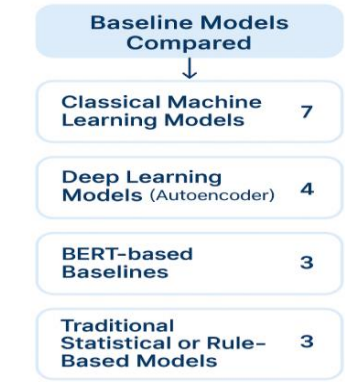

Figure 7. Compared Models

Seven papers evaluated their transformers against classical machine learning methods such as logistic regression, support vector machines, and random forests. These traditional models performed well on structured clinical data and achieved about 60-70% of the modelling strength seen in transformers, highlighting their weakness with temporal patterns and multimodal inputs.

Four studies used deep learning baselines like CNNs, LSTMs, GRUs, and autoencoders, which are naturally better at handling sequential features. They still faced limitations in capturing long-range dependencies across multiple patient visits.

Another set of three studies compared their transformer architectures to earlier transformer-based models such as BERT, BEHRT, or RoBERTa. These models showed that recent models gain additional advantages from domain-specific adaptation and improved sequence encoding.

Three studies also included rule-based or statistical baselines such as simple heuristics or TF-IDF clustering. These methods showed the lowest performance.

H. Discussions

This section integrates fourteen empirical studies. It connects the findings from these studies to trends in clinical machine learning. Architectures in the studies were examined based on their advantages, application scope, architectural evolution, comparative performance, and limitations. [1] [2]

RQ1. Architectural Designs, Modules, and Submodules

The studies reveal that the architectural designs are diverse, reflecting adaptation to a wide range of biomedical data challenges. Semantic understanding enables encoder-only models to dominate textual [21]. Hybrid CNN–Transformer frameworks combine convolutional local feature extraction with attention-based sequence reasoning, as seen in CT-PASMR and LNet Transformer [1], [32].

Architectural competitiveness across transformer models is influenced by design choices such as input representation, attention formulation, and task-specific heads. Encoder-only models are mostly used for textual and phenotyping tasks, while hybrid CNN-Transformer and temporal transformers are better suited for longitudinal and multimodal data [23],[28]. These differences explain performance variability across applications rather than model superiority alone.

1) Hybrid DL + Transformer Architectures

This architecture is a hybrid method of a DL method and transformer-based methods. Studies [1], [29], [38], [40] Demonstrate this hybrid architecture by utilising DL methods and self-attention modules within the same model. DL models are effective in capturing local relationships and in enhancing the ability to fit events in short-term dependencies [1], [29]. Attention mechanisms are effective in capturing global dependencies within longitudinal data [39]. The integration of these models enhances their competitiveness, enabling them to capture both local and long-term dependencies in sequential EHRs and reduce the risk of overfitting.

[1], [29] in the UTHP and CT-PASMR model, they used an RNN module, a CNN module, and a self-attention mechanism. CNN was designed to improve local perception in the position-wise feed-forward branch. At the same time, RNN was used to constrain fitting in temporal

data, and the attention heads were trained to learn global contexts and dependencies. [30] identified a weakness of losing temporal information associated with self-attention heads due to their permutation-invariant nature. Therefore, the TimelyGPT model employed RNN for the task.

2) Pretrained Bio-MED transformers

Pretraining is the process of giving a model broad, general knowledge before refining it for a specific task[10], [41]. This makes pretrained transformers more effective than models built from scratch, because they already understand language structure and contextual cues [42]. BioBERT's exposure to large biomedical text collections enabled it to recognise gene names, disease mentions, and specialised scientific expressions with little additional training [31].

[33], [36] has a similar pattern, where Bio-Clinical-BERT benefits from both general BERT pretraining and additional clinical-domain adaptation. This broad foundation allows the model to interpret short and messy triage notes more effectively than simpler models like BOW-LR, W2V-BiLSTM, or XGBoost.

Bio-Clinical-BERT showed that a model's reliability and understanding of real clinical language are increased by pretraining [36]. This underlines that pretrained models give a richer representation, stronger handling of complex biomedical terminology, and better generalisability [36], [39]. These marks pretraining as important to realise great results in tasks ranging from gene-disease curation to clinical risk prediction.

3) Temporal / Event-Driven Transformers

Temporal data captures changes in health and offers insights into disease progression and treatment [43]. It incorporates EHRs' time-series data and longitudinal data. Study [30] and [33] used masked language modelling and survival modules technique to learn the meaning of temporal data in the context they appear, and a cross-reconstruction transformer to learn temporal classification. Robustness with these techniques gives the models the ability to predict time-series data.

[37] incorporated attention heads, to extract features from input data, a classifier layer to make predictions and a loss function designed to handle missing data. The robust design made it effective for sequential data.

4) Transformer + Graph Modules

A graph is a data structure. It models a set of objects and their relationships, which gives it a great expressive power in ML [44]. In the germline knowledge-graph study, BioBERT's domain-adapted pretraining allows it to recognise complex gene and disease terminology across thousands of abstracts with minimal fine-tuning [31]. This gives the system a clear edge over ontology-only or statistical approaches, enabling more accurate entity extraction before the normalisation stage.

The medication-recommendation model shows a similar benefit. By combining CNN layers with transformer components, the model effectively interpreted a patient's longitudinal EHR history. Graph attention networks in a model learn drug occurrences and different drug-drug interactions, using a joint-loss function. Joint-loss function enables the model to produce more personalised and clinically safer medical predictions and recommendations.

RQ2. Applications of Transformer Architectures

Transformers are flexible, leading to their application across various biomedical domains [21], [23]. They are mainly used for disease prediction and patient stratification. However, they are expanding into omics integration, text mining, and phenotype discovery [20], [22], [45]. Studies leveraging EHR sequences established transformers' ability to encode irregular temporal events, which is necessary for learning disease trajectories [2], [34]. This is a result of the non-Markovian nature of attention, which allows learning dependencies across longitudinal data [29].

From this research's studies, the models were developed with different focus areas.

1) Disease Prediction

Disease prediction remains one of the most impactful uses of DL models in healthcare [11], [46]. Transformer models capture rich patterns across clinical notes, patient histories, and other complex data [33], [39]. They can identify early warning signals that traditional methods often overlook [20], [21]. Utilising the self-attention, they recognise small relationships between symptoms, diagnoses, medications, and past events. With the relationships, they gain a strong advantage in spotting patients at risk [1], [36], [47].

Transformers work well when EHRs or omics data are messy, incomplete, or longitudinal. It learns these complexities to assist clinicians in making predictions earlier and accurately [30], [35].

2) Progression Analysis

Progression analysis monitors how diseases evolve, changes in symptoms, how new conditions develop, and how a patient's overall condition changes [3], [5]. This allows clinicians to perceive how one clinical state leads to another, rather than treating each hospital visit as a separate moment [6], [7].

Transformer-based patient embeddings discovered progression pathways within diseases like colorectal cancer and lupus. Using patient vectors helps review how many diseases in a single phenotype differ [7]. These pathways showed differences in long-term comorbidity burdens and mortality risks.

Transformer models identify early signs of patient trajectories long before diagnosis [34]. They uncover subtypes with distinct risk levels, hospitalisation patterns,

and medication needs [6]. [34] uses contrastive learning to maximise analysis, demonstrating how transformers support a more dynamic and realistic view of disease evolution.

3) Medication Recommendation

Over the years, transformer models have been increasingly used to improve precision medical care [17]. The models propose medications tailored specifically for a patient's condition and genetic variations. Models also recommend safe and accurate prediction of medication needed by a patient, after analysing the condition and drug-drug interactions [1].

TransAMR is a strong example. It uses feature selection algorithms and an ID transformer for pattern recognition in antibiotic use and improves prescription practices. The features are integrated with gradient-based XAI pipelines to interpret insights and recommendations [37].

4) Phenotyping

Phenotyping detects clinical patterns, disease subgroups, or patient characteristics from biomedical data [48]. Earlier methods mostly relied on expert-written rules, but these approaches struggled with incomplete data [49]. To handle these drawbacks, transformer systems learn high-dimensional representations from EHR sequences and clinical notes [50].

[2] revealed clear phenotypic clusters in diseases such as colorectal cancer and lupus. It highlights differences in long-term comorbidity burdens and outcomes [2]. The models can identify disease diversity far earlier than classical approaches [34].

These results drive towards precision medicine, which emphasises deep phenotyping [50]. Transformer-based phenotyping models create more detailed, stable, and have clinically useful representations. They record complex, multi-variable relationships over time [50].

5) Knowledge Extraction

Knowledge extraction converts unstructured biomedical information into clear, structured knowledge [49]. In practice, this step is important in creating a strong predictive model [51]. Rule-based systems or simple statistical techniques had challenges handling complex and specialised language in biomedical text [52]. These challenges were solved by self-attention in transformers [1].

Transformers can understand context. The ability allowed for distinguishing between ambiguous gene symbols and detecting associations, even when they were only implied [53]. This led to a more comprehensive and clinically useful knowledge graph, demonstrating how transformer-driven extraction directly enhances biomedical knowledge curation.

TimelyGPT extends this finding beyond text by extracting structured knowledge from continuous and irregular clinical time-series data [30]. Through extrapolatable xPos embeddings, recurrent attention, and temporal convolutions, the model captures long-term clinical trends and hidden diagnostic patterns that traditional methods overlook.

6) Sequence / Event Modelling

Sequence or event modelling helps comprehend how clinical events unfold over time [37]. Modern transformers perform better at sequence modelling. Their attention mechanisms capture irregular short and long-range patterns. The TransAMR model utilised this method to learn the relationships within antibiotic prescribing sequences [37]. The Universal Transformer Hawkes Process can model the timing and effect of clinical events better than RNN-based methods [29]. The models give a clearer picture of complex clinical directions. They also support more reliable forecasting and decision-making.

Despite this research emphasis on EHR–omics integration, empirical evidence remains heavily skewed toward EHR-only applications. Omics-focused studies were fewer, relied on smaller cohorts, and often prioritised predictive accuracy over biological interpretability. This imbalance highlights a critical disconnect between the theoretical promise of precision medicine and current implementation practices, suggesting that multimodal transformer research is still at an early, exploratory stage.

RQ3. Advantages of Using Transformer Models

Self-attention enables transformer systems to weigh relationships in sequences [33]. This is important in sequence patient data, which requires temporal continuity[49], [54].

Transformers' non-sequential tokenisation and parallel processing make them ideal for heterogeneous data [23], [46]. Their mathematical structure directly aligns with healthcare data characteristics: high dimensionality, contextual dependency, and multimodal complexity [37], [50]. The following is an exploration into the advantages, informed by collective evidence from this research's studies.

1) Captures long-term and sequential patterns in EHR data

Understanding and capturing longitudinal contexts greatly improves the predictive performance of clinical. Unlike traditional DL methods, the transformer's attention mechanism does not suffer from the vanishing gradient effect [1]. DL methods lose information the further they go back in a sequence, making them less reliable for long-term temporal modelling. [1] combines CNN and attention mechanisms, giving the CT-PASMR the ability to record both local and long-range patterns. Also, transformers' bi-directional modelling and parallel processing allow them to learn relationships between distant clinical events without relying on recurrent memory [29]. The models simultaneously draw context from earlier and later events

[40]. This gives a richer and more complete representation of a patient's health.
With these capabilities, the models can monitor disease progression accurately and improve clinical prediction expertise, which are crucial in healthcare

2) Enhances interpretability and clinical insight

Aside from accuracy, models need to be interpretable for them to be implemented in healthcare [10]. They should explain the reasons behind the predictions. Clinicians require transparency to validate recommendations, assess safety and ensure alignment with medical reasoning [39].
Interpretability approaches varied across studies, ranging from attention-weight visualisation and gradient-based attribution to integrated gradients and task-specific explanation layers.Attention visualisation is the most common mechanism, providing coarse temporal saliency but limited causal insight. Models such as TransAMR and CT-PASMR incorporated explicit XAI modules [38]. The modules highlight influential clinical events, while architectures use only attention mechanisms, which do not always translate into clinically meaningful explanations [39]. Post hoc methods (e.g., feature attribution over codes and labs) improve local interpretability but are sensitive to missingness and code sparsity. Targeted designs that embed causal constraints or task-specific rationale layers provide more clinically meaningful explanations but incur higher computational costs. The evidence, therefore, favours pairing temporal attention with task-aware attribution for deployment-grade interpretability. This underscores the absence of standard interpretability practices in transformer-based healthcare models.

3) Cross-domain learning or Generalisation

Generalisation of models is a limiting factor for models in healthcare. It measures how well a model performs on patient representations different from those on which the models are trained [33], [38]. This can be data from a single demographic profile or disease type, on which the model will capture patterns too specific to that environment [39].
Transformer models can improve cross-domain learning through richer representations, multimodal learning and self-attention mechanisms, but they still inherit biases from the data [32].

However, models such as RarePT have shown generalizability enhanced by their masked language modelling feature [35]. It is robust across races, ethnic groups and hospitals.

4) Improves text understanding and extraction of clinical meaning

Clinical notes and biomedical literature contain abbreviations, shifting terminology, and complex phrasing that rule-based methods struggle to interpret [39]. With contextual embeddings, transformers overcome this challenge [38]. The feature enables the extraction of symptoms, diagnoses, and relationships with high accuracy. [31] highlights this strength. It demonstrated how BioBERT identifies gene names, disease terms, and subtle relational cues across more than 11,000 abstracts.
These advantages overlap into real clinical environments. Bio-Clinical-BERT and [39] showed strong text understanding, especially when analysing short and noisy triage notes.
Transformers offer a deeper, more context-aware interpretation of clinical language, making them far more reliable in real healthcare [36].

RQ4. Addressing the Limitations of Transformer Architectures

Research is gradually mitigating known transformer limitations, such as primarily data dependency, generalisability, computational overhead, and interpretability. Transformers require large labelled datasets, which are rare in healthcare due to privacy and heterogeneity [52], [56].
Most reviewed studies prioritised EHR data, with limited empirical fusion of omics modalities. Challenges such as data synchronisation, heterogeneous feature spaces and high dimensionality continue to constrain multimodal transformer development, indicating a significant gap in current research.

1) Limited dataset scope

Healthcare models are trained on datasets with a limited scope [57]. The data is often from a single hospital, region, or demographic group, which ties patterns learnt to that environment [1], [6]. When these models are applied to new patients with different ethnicities or documentation styles, their accuracy often drops[1].
The systems adopt biases found in their training data[1], [30]. Datasets with diverse data can improve generalizability [2]. Even so, limited dataset diversity remains a major problem for deploying clinical AI in real-world settings.

2) Dependence on EHR-only or single-modality data

Most healthcare AI models rely on EHRs. However, EHRs are a small part of a patient's overall health picture. They are non-inclusive of images, omics, lab tests, or physiological monitoring data. This narrows their patient representations [58], [59].
Several of the reviewed transformer models show this limitation. The CT-PASMR model depends entirely on structured EHR and medication data, neglecting omics data [1]. Transformer Patient Embedding model and the Bio-Clinical-BERT research noted that the models relied on patient histories or free-text alone, missing physiological or imaging signals that could produce stronger predictions [2], [36]. . The models overlook key elements, such as understanding biomarkers and mechanisms needed for

precision medicine [60]. Without diverse data integration, AI systems risk remaining narrow, incomplete, and less effective for real-world care.

3) Data quality issues

EHRs often contain missing diagnoses, incomplete data, unrecorded medications, and inconsistent visit histories. Misspelt clinical terms, vague symptom descriptions, or outdated information weaken model performance [58]. These issues distort learned patterns [24], [45].

[36] highlights how triage notes are often not consistent or complete, making it difficult for models to extract reliable features. [1], [2] also depend on the patient's records, where key clinical events may be missing or underdocumented.

Poor data quality reduces accuracy and generalizability. It increases the chances of unreliable or unsafe clinical predictions [58]. Strategies such as imputation, data cleaning, and multimodal integration can reduce noise in EHRs.

4) Limited interpretability or transparency

Interpretability remains a challenge in AI models. Although transformer architectures have improved predictive accuracy and contextual understanding of data, they are referred to as black boxes [29]. Clinicians require explanations of insights and recommendations given by the model [37].

Many transformer models provide attention scores or gradient-based attribution, but the models do not translate into clinically meaningful explanations [34]. The insights generated by the model remain opaque, limiting adoption in practice. Until AI systems can offer reliable and clinically grounded explanations, their use in decision-making remains restricted.

RQ5. Comparative Performance with Baseline Models

Transformer models consistently outperform CNN, RNN, and classical ML algorithms because they can generalise beyond local patterns and exploit contextual associations [45]. Their superiority across tasks in this review [1], [2], [61] aligns with benchmarking studies showing transformers' scalability and expressive capacity [49]. For instance, CT-PASMR achieved higher accuracy and interpretability than recurrent networks by integrating convolutional filters for local dependencies with self-attention for global sequence modelling [1]. Bio-Clinical-BERT improved hospital-admission prediction by 608% in AUROC. The following sub-section will provide an in-depth discussion on transformers and other models in research.

1) Classical Machine Learning Models

Classical ML models like Logistic Regression, XGBoost, and SVM were commonly used as baselines in several of the reviewed studies [38], [39]. Findings from the Bio-Clinical-BERT show that LR-TF-IDF models achieved AUC scores of 0.81-0.84, while transformer models slightly outperformed them with AUC values of 0.82-0.85 [36]. In the Bio-Clinical-BERT triage-note study, classical models such as Logistic Regression and XGBoost achieved AUC values between 0.76 and 0.84, performing reasonably well on structured or shallow text features but still slightly below the transformer model's 0.82-0.85 range [36].

However, the study also revealed that transformers consistently outperform classical baselines as tasks become more complex. For deeper contextual understanding or long-range pattern modelling, transformers show a clear performance gap of 20-30% [1], [32]. Classical ML relies on fixed feature engineering and cannot capture semantic relationships, which limits its predictive ability.

2) Deep Learning Models

CNNs, LSTMs, and GRU networks appeared often as baselines. In the TransAMR system, the 1D-CNN and modified ResNet had challenges with complex datasets. They fall about 10-20% behind of the TransAMR in masked AUC and F1 performance [37]. This proves that traditional deep models can miss longer-range relationships in antibiotic use patterns.

CT-PASMR identified similar trends. Models like the LSTM-based LEAP recorded short-term visit patterns but scored lower on Jaccard, F1, and recall [1]. Adding transformer attention improved performance by roughly 5-10%, highlighting its ability to capture broader patient history.

[2] argues that while the variational autoencoder baseline produced useful embeddings, it could not model patient trajectories over time. Transformer-based embeddings performed about 8-12% better in downstream disease forecasting. Overall, deep learning models provided solid baselines, but transformers consistently delivered stronger results, especially for tasks that require long-range reasoning or richer clinical context.

3) BERT-based Baselines

Domain-specific BERT models have shown clear advantages over the general BERT model in biomedical and clinical tasks [33], [39]. General BERT often struggles with the specialised terms, abbreviations, and gene or disease names that appear in scientific and clinical writing [62]. As a result, models trained on biomedical text consistently outperform standard BERT with named-entity recognition and relation extraction [31], [36].

With BioBERT, domain knowledge helped it identify genes, diseases, and relationships more accurately than TF-IDF and other statistical baselines [31]. Bio-Clinical-BERT also demonstrated stronger performance than general BERT when analysing triage notes and EHR narratives [36]. Domain-specific BERT models provide deeper contextual understanding and more reliable clinical results. This makes them stronger baselines than the original BERT model for healthcare applications.

4) Traditional Statistical or Rule-Based Models

The TF-IDF and K-means approach produces very weak clustering quality, with a Silhouette score of just 0.05 [34]. In contrast, the transformer-based model achieved a Silhouette score of 0.36. This illustrates how statistical methods struggle to capture the complex temporal and multi-comorbidity patterns present in real patient data. Also, TF-IDF clustering and statistical baselines cannot manage synonym variation or the specialised structure of biomedical text. As a result, the models cannot accurately recognise biomedical entities or extract relationships [31].

Rule-based and statistical models are computationally efficient and easy to interpret [63]. However, they have limited flexibility, especially when dealing with large, heterogeneous clinical datasets [64]. [34] highlights the TF-IDF + K-means pipeline produced a low Calinski–Harabasz score of 2,191, compared to the transformer's 23,371 score.

Traditional methods cannot differentiate ambiguities within gene symbols[34]. These findings show that rule-based and statistical approaches consistently underperform compared to transformer-based architectures.

Across the 14 studies, transformer architectures that combine longitudinal self-attention with biomedical pre-training (e.g., BEHRT-family variants) consistently outperform recurrent and classical baselines on disease prediction and phenotyping, particularly where long-range temporal dependencies dominate. However, gains are attenuated on small or single-institution cohorts and in settings with high missingness, indicating sensitivity to data quality and cohort shift. Hybrid designs that integrate temporal attention with task-specific heads (e.g., medication recommendation) show the most reliable improvements, while purely generic encoders exhibit greater variance across tasks.

H. Implications

1) Theoretical Implications

Theoretically, transformers can capture long-range and non-linear relationships in data. The relationships make the models adaptive. The non-Markovian attention mechanism leverages them to model complex and irregular connections between molecular, clinical, and textual data [54]. Older sequential models, such as RNNs fail to achieve this ability. This strength reflects principles from systems biology, which view disease as a network of interconnected omics and clinical factors [46], [65]. This way, the transformer's predictions of disease projections are improved. It also offers a theoretical basis that pushes the field toward truly integrated precision medicine.

2) Practical Implications

Transformer-based models have revolutionised healthcare with new applications of AI. Adding various interpretability aids, such as attention heatmaps and attribution methods, can accelerate their clinical adoption. Interpretability increases the system's transparency and trustworthiness, which supports acceptance among clinicians and regulators [34], [37].

Pretrained models like Bio-Clinical-BERT and task-specific architectures like LNet can be fine-tuned efficiently, lowering computational barriers for use in smaller or data-scarce health systems.

However, high computational costs and unequal data representation still threaten bias, accessibility and reliability in real-world settings [50], [66].

I. Limitations of Study

The study only reviewed papers in four databases, IEEE, PubMed, ACM Digital Library and ScienceDirect. This selection neglects other studies available in other databases but not in the ones chosen for this study. Exclusion of studies with models trained or papers written in languages other than English introduces bias to the study. It misses other architectural designs used for different linguistic data and useful research in other languages. Also, the studies included in the study focused on EHRs or omics data, neglecting other modalities such as medical images and radiology reports. This narrow data scope undermines the full capabilities of transformer models emerging in healthcare.

J. Future Works

Future research should focus on bridging interpretability and generalisation limitations within healthcare AI models. Transformer frameworks should incorporate explainable artificial intelligence. They should also integrate omics and multi-omics data with EHRs. Omics and multi-omics reveal the basic principles of biological functions. Therefore, with EHRs classification and health predictions, models can learn reasons for patient trajectories. Explainability proves AI models' reliability and trustworthiness, which determines their adoption in healthcare settings.

Moreover, the research should train models on diverse modalities. This is to ensure ethnic groups' health and biological patterns are recognised by models. Developing models like RarePT, which can be applied to diverse health care instances regardless of race, ethnicity or hospital, can also increase generalisability.

IV. CONCLUSION

This systematic review demonstrates how transformer architectures are enhancing biological insights through the integration and interpretability of EHRs and omics. They learn longitudinal temporal dependencies. Models such as CT-PASMR, Bio-Clinical-BERT, and LNet Transformer exemplify how self-attention mechanisms can optimise

medication recommendations, improve hospitalisation prediction, and integrate multi-omics signals for disease manifestation.

Despite these advances, studies remain largely EHR-centric. Omics integration remains limited, undermining the understanding of biomarkers, functional pathways and mechanisms. Limited generalisability due to single-institution data or small cohorts, computational inefficiency, missing data, or interpretability challenges remain within models.

Overall, transformers are a game-changer for biomedical data fusion. Their capacity for hierarchical learning with contextual awareness has improved predictive accuracy, interpretability and model trustworthiness.

For clinicians, transformer-based systems offer enhanced risk stratification and decision support but require transparent explanations to support trust and safety. For researchers, findings highlight the need for multimodal datasets, robust external validation, and standardised interpretability evaluation. For developers of clinical decision support systems, scalable architectures and efficient training strategies are essential for real-world deployment. Addressing these challenges will be critical to translating transformer-based models from experimental studies into routine clinical practice.

REFERENCES

[1] F. Ge, X. Yu, X. Li, X. Fan, and Y. Zhao, "Personalized and safe medication recommendation based on convolutional neural network and transformer architecture," *Eng. Appl. Artif. Intell.*, vol. 161, Dec. 2025, doi: 10.1016/j.engappai.2025.112267.

[2] S. X. E. G. J. K. L. W. S. L. W.-Q. W. P. J. D. M. R. Crosslin, "Transformer patient embedding using electronic health records enables patient stratification and progression analysis," *npj Digit. Med.*, 2025, doi: 10.1038/s41746-025-01872-z.

[3] G. Chitra and S. M. Basha, "Systematic Literature Review on Deep Learning Techniques in Electronic Health Records," *2023 4th Int. Conf. Electron. Sustain. Commun. Syst. ICESC 2023 - Proc.*, no. January, pp. 1365–1371, 2023, doi: 10.1109/ICESC57686.2023.10193025.

[4] R. Sibanda, B. Ndlovu, S. Dube, and K. Maguraushe, "Towards Health 4 . 0 : Blockchain-Based Electronic Health Record for Care Coordination," pp. 712–720, 2024, doi: 10.34190/ecie.19.1.2606..

[5] M. S. Safarova and I. J. Kullo, "Using the electronic health record for genomics research," *Curr. Opin. Lipidol.*, vol. 31, no. 2, pp. 85–93, 2020, doi: 10.1097/MOL.0000000000000662.

[6] S. A. Pendergrass and D. C. Crawford, "Using Electronic Health Records To Generate Phenotypes For Research," *Curr. Protoc. Hum. Genet.*, vol. 100, no. 1, pp. 1–28, 2019, doi: 10.1002/cphg.80.

[7] K. Himavamshi, D. Tejaswini, G. Sethi, V. S. A. Devi, P. Pavani, and S. Hariharan, "Electronic Health Record classification and analysis using NLP Techniques," *E3S Web Conf.*, vol. 619, 2025, doi: 10.1051/e3sconf/202561903016.

[8] M. R. K. Goyal, "Integrating omics data for personalized medicine in treating psoriasis," *Med. Chem. Res.*, 2024, doi: 10.1007/s00044-024-03355-4.

[9] Y. Wu, Y. Cheng, X. Wang, J. Fan, and Q. Gao, "Spatial omics: Navigating to the golden era of cancer research," *Clin. Transl. Med.*, vol. 12, no. 1, 2022, doi: 10.1002/ctm2.696.

[10] Y. Ren *et al.*, "COMET: Benchmark for Comprehensive Biological Multi-omics Evaluation Tasks and Language Models," pp. 1–29, 2024, [Online]. Available: http://arxiv.org/abs/2412.10347

[11] A. E. Mohr, C. P. Ortega-Santos, C. M. Whisner, J. Klein-Seetharaman, and P. Jasbi, "Navigating Challenges and Opportunities in Multi-Omics Integration for Personalized Healthcare," *Biomedicines*, vol. 12, no. 7, 2024, doi: 10.3390/biomedicines12071496.

[12] L. Zhao *et al.*, "DeepOmix: A scalable and interpretable multi-omics deep learning framework and application in cancer survival analysis," *Comput. Struct. Biotechnol. J.*, vol. 19, pp. 2719–2725, 2021, doi: 10.1016/j.csbj.2021.04.067.

[13] J. Zhao, Q. Feng, and W.-Q. Wei, "Integration of Omics and Phenotypic Data for Precision Medicine.," *Methods Mol. Biol.*, vol. 2486, pp. 19–35, 2022, doi: 10.1007/978-1-0716-2265-0_2.

[14] M. L. G. G. T. T. Y. J. S. Jia, "Machine learning and multi-omics integration: advancing cardiovascular translational research and clinical practice," *J. Transl. Med.*, 2025, doi: 10.1186/s12967-025-06425-2.

[15] I. C. Udousoro, "Machine Learning: A Review," *Semicond. Sci. Inf. Devices*, vol. 2, no. 2, pp. 5–14, 2020, doi: 10.30564/ssid.v2i2.1931.

[16] S. Hadebe, B. Ndlovu, and K. Maguraushe, "Managing Diabetes Using Machine Learning and Digital Twins," pp. 145–162, 2025, doi: 10.47540/ijias.v5i2.1981.

[17] I. S. Mangkunegara, Purwono, A. Ma'arif, N. Basil, H. M. Marhoon, and A. N. Sharkawy, "Transformer Models in Deep Learning: Foundations, Advances, Challenges and Future Directions," *Bul. Ilm. Sarj. Tek. Elektro*, vol. 7, no. 2, pp. 231–241, 2025, doi: 10.12928/biste.v7i2.13053.

[18] A. Behrouz, P. Zhong, and V. Mirrokni, "Titans: Learning to Memorize at Test Time," pp. 1–27, 2024, [Online]. Available: http://arxiv.org/abs/2501.00663

[19] A. Mohamed, R. AlAleeli, and K. Shaalan, "Advancing Predictive Healthcare: A Systematic Review of Transformer Models in Electronic Health Records," *Computers*, vol. 14, no. 4, pp. 1–28, 2025, doi: 10.3390/computers14040148.

[20] V. A. Batista and A. G. Evsukoff, "Application of Transformers based methods in Electronic Medical Records: A Systematic Literature Review," pp. 1–28, 2023, [Online]. Available: http://arxiv.org/abs/2304.02768

[21] 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. doi: 10.1007/s10462-023-10677-z.

[22] S. Shool, S. Adimi, R. Saboori Amleshi, E. Bitaraf, R. Golpira, and M. Tara, "A systematic review of large language model (LLM) evaluations in clinical medicine," *BMC Med. Inform. Decis. Mak.*, vol. 25, no. 1, 2025, doi: 10.1186/s12911-025-02954-4.

[23] A. Mohamed, R. AlAleeli, and K. Shaalan, "Advancing Predictive Healthcare: A Systematic Review of Transformer Models in Electronic Health Records," *Computers*, vol. 14, no. 4, 2025, doi: 10.3390/computers14040148.

[24] A. Mohamed, R. AlAleeli, and K. Shaalan, "Advancing Predictive Healthcare: A Systematic Review of Transformer Models in Electronic Health Records," Apr. 01, 2025, *Multidisciplinary Digital Publishing Institute (MDPI)*. doi: 10.3390/computers14040148.

[25] M. J. Page *et al.*, "PRISMA 2020 explanation and elaboration: Updated guidance and exemplars for reporting systematic reviews," *BMJ*, vol. 372, 2021, doi: 10.1136/bmj.n160.

[26] T. Dybå and T. Dingsøyr, "Empirical studies of agile software development: A systematic review," *Inf. Softw. Technol.*, vol. 50, no. 9–10, pp. 833–859, 2008, doi: 10.1016/j.infsof.2008.01.006.

[27] B. Kitchenham, "Kitchenham , B .: Guidelines for performing Systematic Literature Reviews in software engineering . EBSE Technical Report EBSE-2007-01 Guidelines for performing Systematic Literature Reviews in Software Engineering," *Icse*,

no. January 2007, pp. 1–57, 2007.
[28] S. Keele, "Guidelines for performing systematic literature reviews in software engineering," *Tech. report, Ver. 2.3 EBSE Tech. Report. EBSE*, no. October, 2007.
[29] L. ning Zhang, J. wei Liu, Z. yan Song, and X. Zuo, "Universal transformer Hawkes process with adaptive recursive iteration," *Eng. Appl. Artif. Intell.*, vol. 105, Oct. 2021, doi: 10.1016/j.engappai.2021.104416.
[30] Z. Song, Q. Lu, H. Xu, H. Zhu, D. Buckeridge, and Y. Li, "TimelyGPT: Extrapolatable Transformer Pre-training for Long-term Time-Series Forecasting in Healthcare," in *Proceedings of the 15th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*, in BCB '24. New York, NY, USA: Association for Computing Machinery, 2024. doi: 10.1145/3698587.3701364.
[31] A. D. Diaz Gonzalez, K. S. Hughes, S. Yue, and S. T. Hayes, "Applying BioBERT to Extract Germline Gene-Disease Associations for Building a Knowledge Graph from the Biomedical Literature," in *Proceedings of the 2023 7th International Conference on Information System and Data Mining*, in ICISDM '23. New York, NY, USA: Association for Computing Machinery, 2023, pp. 37–42. doi: 10.1145/3603765.3603771.
[32] R. Miao *et al.*, "An Integrated Multi-omics prediction model for stroke recurrence based on Lnet transformer layer and dynamic weighting mechanism," *Comput. Biol. Med.*, vol. 179, Sep. 2024, doi: 10.1016/j.compbiomed.2024.108823.
[33] 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, Jun. 2025, doi: 10.1016/j.landig.2025.03.005.
[34] Z. Fan, M. Mamouei, Y. Li, S. Rao, and K. Rahimi, "Identification of heart failure subtypes using transformer-based deep learning modelling: a population-based study of 379,108 individuals," *eBioMedicine*, vol. 114, Apr. 2025, doi: 10.1016/j.ebiom.2025.105657.
[35] D. M. Jordan, H. M. T. Vy, and R. Do, "A deep learning transformer model predicts high rates of undiagnosed rare disease in large electronic health systems," *medRxiv*, p. 2023.12.21.23300393, Dec. 2023, doi: 10.1101/2023.12.21.23300393.
[36] D. Patel *et al.*, "Traditional Machine Learning, Deep Learning, and BERT (Large Language Model) Approaches for Predicting Hospitalizations From Nurse Triage Notes: Comparative Evaluation of Resource Management," *JMIR AI*, vol. 3, no. 1, 2024, doi: 10.2196/52190.
[37] M. Tharmakulasingam, W. Wang, M. Kerby, R. La Ragione, and A. Fernando, "TransAMR: An Interpretable Transformer Model for Accurate Prediction of Antimicrobial Resistance Using Antibiotic Administration Data," *IEEE Access*, vol. 11, pp. 75337–75350, 2023, doi: 10.1109/ACCESS.2023.3296221.
[38] X. Zhou *et al.*, "CMACF: Transformer-based cross-modal attention cross-fusion model for systemic lupus erythematosus diagnosis combining Raman spectroscopy, FTIR spectroscopy, and metabolomics," *Inf. Process. Manag.*, vol. 61, no. 6, Nov. 2024, doi: 10.1016/j.ipm.2024.103804.
[39] X. Luo *et al.*, "Applying interpretable deep learning models to identify chronic cough patients using EHR data," *Comput. Methods Programs Biomed.*, vol. 210, Oct. 2021, doi: 10.1016/j.cmpb.2021.106395.
[40] S. Naveed, M. Husnain, and N. Alsubaie, "HybridDLDR: A hybrid deep learning-based drug resistance prediction system of Glioblastoma (GBM) using molecular descriptors and gene expression data," *Comput. Methods Programs Biomed.*, vol. 270, Oct. 2025, doi: 10.1016/j.cmpb.2025.108913.
[41] X. Wang *et al.*, "Hierarchical Pretraining on Multimodal Electronic Health Records," *EMNLP 2023 - 2023 Conf. Empir. Methods Nat. Lang. Process. Proc.*, pp. 2839–2852, 2023, doi: 10.18653/v1/2023.emnlp-main.171.
[42] M. McDermott *et al.*, "A comprehensive EHR timeseries pre-training benchmark," *ACM CHIL 2021 - Proc. 2021 ACM Conf. Heal. Inference, Learn.*, pp. 257–278, 2021, doi: 10.1145/3450439.3451877.
[43] A. Patharkar, F. Cai, F. Al-Hindawi, and T. Wu, "Predictive modeling of biomedical temporal data in healthcare applications: review and future directions," *Front. Physiol.*, vol. 15, no. October, pp. 1–24, 2024, doi: 10.3389/fphys.2024.1386760.
[44] J. Zhou *et al.*, "Graph neural networks: A review of methods and applications," *AI Open*, vol. 1, no. September 2020, pp. 57–81, 2020, doi: 10.1016/j.aiopen.2021.01.001.
[45] S. R. Choi and M. Lee, "Transformer Architecture and Attention Mechanisms in Genome Data Analysis: A Comprehensive Review," *Biology (Basel).*, vol. 12, no. 7, 2023, doi: 10.3390/biology12071033.
[46] L. Tong *et al.*, "Integrating Multi-Omics Data With EHR for Precision Medicine Using Advanced Artificial Intelligence," *IEEE Rev. Biomed. Eng.*, vol. 17, pp. 80–97, 2024, doi: 10.1109/RBME.2023.3324264.
[47] S. Madan, M. Lentzen, J. Brandt, D. Rueckert, M. Hofmann-Apitius, and H. Fröhlich, "Transformer models in biomedicine," *BMC Med. Informatics Decis. Mak. 2024 241*, vol. 24, no. 1, pp. 1–22, Jul. 2024, doi: 10.1186/S12911-024-02600-5.
[48] J. Li *et al.*, "An integrated pipeline for prediction of Clostridioides difficile infection.," *Sci. Rep.*, vol. 13, no. 1, p. 16532, Oct. 2023, doi: 10.1038/s41598-023-41753-7.
[49] B. Shickel, P. J. Tighe, A. Bihorac, and P. Rashidi, "Deep EHR: A Survey of Recent Advances in Deep Learning Techniques for Electronic Health Record (EHR) Analysis," *IEEE J. Biomed. Heal. Informatics*, vol. 22, no. 5, pp. 1589–1604, 2018, doi: 10.1109/JBHI.2017.2767063.
[50] A. Rajkomar, J. Dean, and I. Kohane, "Machine Learning in Medicine," *N. Engl. J. Med.*, vol. 380, no. 14, pp. 1347–1358, 2019, doi: 10.1056/nejmra1814259.
[51] Y. Wang *et al.*, "Clinical information extraction applications: A literature review," *J. Biomed. Inform.*, vol. 77, no. November 2017, pp. 34–49, 2018, doi: 10.1016/j.jbi.2017.11.011.
[52] A. Vaid *et al.*, "Using Deep-Learning Algorithms to Simultaneously Identify Right and Left Ventricular Dysfunction From the Electrocardiogram," *JACC Cardiovasc. Imaging*, vol. 15, no. 3, pp. 395–410, Mar. 2022, doi: 10.1016/j.jcmg.2021.08.004.
[53] J. T. VanSchaik, P. Jain, A. Rajapuri, B. Cheriyan, T. P. Thyvalikakath, and S. Chakraborty, "Using transfer learning-based causality extraction to mine latent factors for Sjögren's syndrome from biomedical literature," *Heliyon*, vol. 9, no. 9, Sep. 2023, doi: 10.1016/j.heliyon.2023.e19265.
[54] J. U. Ashish Vaswani, Noam Shazeer, Niki Parmar, "Attention Is All You Need Ashish," *IEEE Ind. Appl. Mag.*, vol. 8, no. 1, pp. 8–15, 2017, doi: 10.1109/2943.974352.
[55] 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.
[56] A. Vaid *et al.*, "Federated learning of electronic health records to improve mortality prediction in hospitalized patients with COVID-19: Machine learning approach," *JMIR Med. Informatics*, vol. 9, no. 1, Jan. 2021, doi: 10.2196/24207.
[57] T. Wang *et al.*, "MOGONET integrates multi-omics data using graph convolutional networks allowing patient classification and biomarker identification," *Nat. Commun.*, vol. 12, no. 1, pp. 1–13, 2021, doi: 10.1038/s41467-021-23774-w.
[58] T. Ching *et al.*, *Opportunities and obstacles for deep learning in biology and medicine*, vol. 15, no. 141. 2018. doi: 10.1098/rsif.2017.0387.
[59] R. Miotto, L. Li, B. A. Kidd, and J. T. Dudley, "Deep Patient: An Unsupervised Representation to Predict the Future of Patients from the Electronic Health Records," *Sci. Rep.*, vol. 6, no. January, pp. 1–10, 2016, doi: 10.1038/srep26094.

[60] S. J. M. A. E. S. M. R. B. D. R. K. G.-H. S. J. T. S. A. S. F. J. W. M. S. S. A. G. K. S. Aghaeepour, "Author Correction: A machine learning approach to leveraging electronic health records for enhanced omics analysis," *Nat. Mach. Intell.*, 2025, doi: 10.1038/s42256-025-01021-x.

[61] Z. M. D. H. P. M. Kim, "Multi-omics integration in the age of million single-cell data," *Nat. Rev. Nephrol.*, 2021, doi: 10.1038/s41581-021-00463-x.

[62] J. Lee *et al.*, "BioBERT: A pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics*, vol. 36, no. 4, pp. 1234–1240, 2020, doi: 10.1093/bioinformatics/btz682.

[63] A. Alnattah, M. Jajroudi, S. A. N. Fadafen, M. N. Manzari, and S. Eslami, "Artificial Intelligence in Clinical Decision-Making: A Scoping Review of Rule-Based Systems and Their Applications in Medicine," *Cureus*, vol. 17, no. 8, 2025, doi: 10.7759/cureus.91333.

[64] B. Silva, F. Hak, T. Guimaraes, M. Manuel, and M. F. Santos, "Rule-based System for Effective Clinical Decision Support," *Procedia Comput. Sci.*, vol. 220, no. 2019, pp. 880–885, 2023, doi: 10.1016/j.procs.2023.03.119.

[65] Y. Hasin, M. Seldin, and A. Lusis, "Multi-omics approaches to disease," *Genome Biol.*, vol. 18, no. 1, pp. 1–15, 2017, doi: 10.1186/s13059-017-1215-1.

[66] V. Sanh *et al.*, "Multitask Prompted Training Enables Zero-Shot Task Generalization," *ICLR 2022 - 10th Int. Conf. Learn. Represent.*, 2022.