Artificial Intelligence In Medicine 158 (2024) 102999

Contents lists available at ScienceDirect

Artificial Intelligence In Medicine

journal homepage: www.elsevier.com/locate/artmed

A systematic review of networks for prognostic prediction of health outcomes and diagnostic prediction of health conditions within Electronic Health Records

Zoe Hancox [a,*], Allan Pang [a,b], Philip G. Conaghan [c,d], Sarah R. Kingsbury [c,d], Andrew Clegg [a], Samuel D. Relton [a]

[a] University of Leeds, Leeds, United Kingdom
[b] Royal Centre for Defence Medicine, Research & Clinical Innovation (RCI), ICT Centre, Vincent Drive, Birmingham, United Kingdom
[c] Leeds Institute of Rheumatic and Musculoskeletal Medicine, University of Leeds, Leeds, United Kingdom
[d] NIHR Leeds Biomedical Research Centre, United Kingdom

ARTICLE INFO

Keywords:
Graphs
Networks
Electronic health records
Prediction
Machine learning

ABSTRACT

Background and objective Using graph theory, Electronic Health Records (EHRs) can be represented graphically to exploit the relational dependencies of the multiple information formats to improve Machine Learning (ML) prediction models. In this systematic qualitative review, we explore the question: How are graphs used on EHRs, to predict diagnosis and health outcomes?

Methodology The search strategy identified studies that used patient-level graph representations of EHRs to utilise ML to predict health outcomes and diagnoses. We conducted our search on MEDLINE, Web of Science and Scopus.

Results 832 studies were identified by the search strategy, of which 27 studies were selected for data extraction. Following data extraction, 18 studies used ML with patient-level graph-based representations of EHRs to predict health outcomes and diagnoses. Models ranged from traditional ML to neural network-based models. MIMIC-III was the most used dataset (n = 6, where n is the number of occurrences), followed by National Health Insurance Research Database (NHIRD) (n = 4) and eICU Collaborative Research Database (eICU) (n = 4). The most predicted health outcomes were mortality (n = 9; 21%), hospital readmission (n = 9; 21%), and treatment success (n = 4; 9%). Model performances ranged across outcomes, mortality prediction (Area Under the Receiver Operating Characteristic (AUROC): 72.1 - 91.6; Area Under Precision-Recall Curve (AUPRC): 34.8 - 81.3) and readmission prediction (AUROC: 63.7 - 85.8; AUPRC 39.86 - 84.7). Only one paper had a low Risk of Bias (RoB) that applied to our research question (4%).

Conclusion Graph-based representations using EHRs, for individual health outcomes and diagnoses requires further research before we can see the results applied clinically. The use of graph representations appears to improve EHR representation and predictive performance compared to baseline ML methods in multiple fields of medicine.

Contents

* Corresponding author.
E-mail addresses: Z.L.Hancox@leeds.ac.uk (Z. Hancox), allan.pang@nhs.net (A. Pang), S.D.Relton@leeds.ac.uk (S.D. Relton).

https://doi.org/10.1016/j.artmed.2024.102999
Received 5 June 2023; Received in revised form 1 October 2024; Accepted 16 October 2024
Available online 23 October 2024

1. Introduction

Improvements in medical advances with increasing complexities of patient treatment pathways and ageing demographics have increased pressure on healthcare services. Implementing predictive algorithms into healthcare settings can reduce the cognitive burden for clinicians whilst reducing patient wait time for care [1].

EHRs are used within clinical practice to document and store patient data during clinical encounters. EHRs contain a wealth of patient data, including health events, symptoms, laboratory investigations, and diagnoses [2]. Most clinical prediction models use summary data, such as EHR codes, which alone loses the inherent structure and temporality of the data.

Graph theory utilises network structures and uses mathematics to observe patterns and structures within data. In discrete mathematics, a graph $G = (V, E)$ is defined as a series of nodes V connected via edges E to represent relationships between nodes [3,4].

Graphs can be used to model EHR data to maintain structural features, temporality, and comorbidities which can be used to predict patient outcomes with ML. Earlier predictions of health outcomes may allow preventive interventions to be carried out (e.g., physiotherapy, medication), which can reduce the impact of healthcare utilisation, lessen patient suffering, prevent conditions from worsening, and reduce future healthcare utilisation.

Graph representations for EHRs are becoming increasingly popular. Social network analysis methods can be used to find disease progression by finding similarities between patients and their outcome trajectories [5,6]. Patients can be clustered based on graphical EHR representations to make diagnoses [7]. Increasingly Deep Learning (DL) methods, such as Convolutional Neural Networks (CNNs) and Graph Neural Networks (GNNs), are being used to find important features and patterns in an individual patient EHR to predict patient prognoses [8,9].

This systematic literature review follows the PRISMA guidelines for comprehensive literature search and selection. It also uses the PROBAST criteria to assess the RoB and quality of papers to investigate the utilisation of graphs in healthcare for patient-level EHR representations and health predictions. The research question guiding this review is:

How are graphs being used on EHRs to predict diagnosis and health outcomes?

To address this broader question, we answer the following sub-questions:

(a) What graph approaches are researchers taking to predict these health outcomes?
(b) How do these approaches compare to other ML, Artificial Intelligence (AI), and statistical models?
(c) How are nodes and edges being utilised to perform these tasks?
(d) How do these graph approaches compare to each other?

Outcomes from these studies are highly heterogeneous making a meta-analysis inappropriate. Instead, results are presented as a narrative synthesis, comparison, and discussion of studies. To the best of our knowledge, this systematic review is one of the first to analyse studies that have used graph techniques to represent EHR information for health prediction and diagnosis.

2. Related work

Schrodt et al. review on graph representations of patient data (extraction date: 20-Mar-2018) is the only systematic literature review paper identified that focuses on graph representations of patient data [10]. Contrasting our review that examines how graph representations are used for prediction, Schrodt's review of 11 articles examines how graphs were used to represent EHR of individual patients.

There are five systematic reviews that focus on ML based prediction tasks using EHRs as input data. These do not focus solely on graphical representations. Two papers explored DL models using EHR, but neither retrieved any graph-based models [11,12].

Si et al. review on deep representation learning of EHR (extraction period: 2015–2019) identified 49 papers [13]. They discussed graph-based patient representation, models such as GNNs, and highlighted various works (n = 8). Three references in their paper also match our included Refs. [7,13,14]. Si et al. suggested that future work will involve harnessing the complex features found in EHRs, improving reproducibility and transparency.

Liu et al. review concentrated on representation learning of EHRs and suggested categorising these methods into statistical, knowledge-based, and graph learning methods [2]. There are four papers in Liu's review that appeared in our search [14–17]. Liu suggested that graphs are a practical way to represent EHRs that maintains the structural, temporal, and semantic relationships, which is not possible with other methods.

Hossain et al. review of 36 papers (extraction date: Aug-2018) explored the use of EHR data for disease prediction [18]. One of their papers was included in our search [16]. This review found that different MLs methods worked best for different clinical settings. Graph-based methods appeared to work best for Diabetes mellitus (DM), and

Hossain suggested that graph representations enable the relationships between healthcare data to be structured, enabling an understanding of connections that otherwise might be difficult to observe.

3. Systematic review methods

This systematic review follows the 2015 PRISMA protocols [19]. The completed checklist can be found in Appendix A.8. Our review is registered on PROSPERO with the protocol registration number CRD42022315782.

3.1. Search strategy

Our search strategy involves combining synonyms for "Graphs" and "Electronic Health Records". Asterisk wildcards were applied to "Prognostic", "Diagnostic", and "Prediction" to expand the search. Queries targeted abstracts, titles, and keywords. Studies within this graph/network domain were evaluated to determine if other terms cover the same concept. The search, conducted on February 27, 2023, covered MEDLINE, Scopus, and Web of Science databases. A forward citation search of review articles identified at the abstract title screening stage was conducted. The complete search terms are in Appendix A.1.

3.2. Inclusion criteria

In our review, "graph" specifically refers to graph theory, representing information in a network of nodes and edges. The common meaning of charts/visualisations was excluded. Included studies constructed graphs directly from individual patient-level EHR data, excluding those using aggregated population data. To assess the effectiveness of graph representation in ML-based predictions, we only considered primary research studies that described at least one ML prediction task using EHR graph representation.

Outcomes are defined as diagnostic prediction of a health condition (e.g. Heart Failure (HF) or cancer) or prognostic prediction of a health-related outcome (e.g. mortality, readmission risk, or treatment success). Studies that predicted multi-class outcomes with over ten possible labels were excluded, as statistical reporting of these models is insufficient.

Graph-based learning in healthcare gained recent attention [1]. Considering the impact of hardware availability on DL progression, we focus on papers published between 2002 and 2023 to align with the release of Torch, a popular ML library framework in 2002.

We excluded grey literature (theses, dissertations, non-peer reviewed pre-prints, and online repositories), only including full-text papers written in English or with an English translation.

3.3. Article selection

ZH and AP independently conducted title/abstract and full-text screening stages, reaching a consensus on selection at each stage. Disagreements were resolved by SR, the third reviewer, and Rayyan's online software tool was employed in this process [20].

4. Data extraction methods

We employed two assessment frameworks, evaluated RoB, and identified characteristics to ensure a reproducible study assessment. ZH and AP independently conducted both assessments for each study.

4.1. Risk of bias (RoB)

We evaluated the RoB using the PROBAST tool, a framework for assessing the quality of methodologies, including RoB and applicability, in primary studies developing prediction models for diagnosis and prognosis [21].

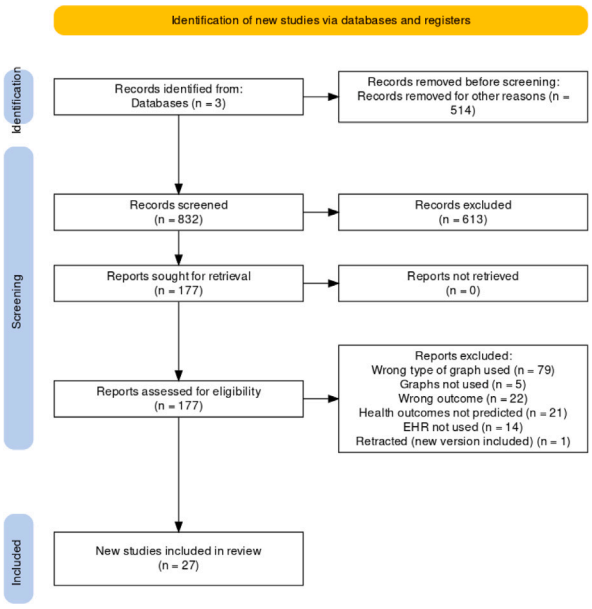

Fig. 1. PRISMA flow diagram illustrating the search strategy. Figure generated with the tool from [28].

PROBAST, developed through a consensus process with 20 signalling questions across four domains (participants, predictors, outcome, and analysis), assigns a RoB score of High, Low, or Unknown to each domain based on signalling questions. The overall ROB is determined by the worst domain score (i.e., an overall low RoB requires every domain to score low) [21]. Reviewers reached a consensus on RoB at the domain level, and a qualitative analysis was conducted for overall and domain-specific RoB across all studies. We include papers with high RoB to highlight ongoing work in the field, focusing on the primary health outcome when multiple models are developed in a study.

4.2. Study characteristics

We utilised an adapted CHecklist for critical Appraisal and data extraction for systematic Reviews of prediction Modelling Studies (CHARMS) for extracting study characteristics, originally designed for primary studies on diagnostic or prognostic prediction models [22]. Qualitative analysis of the data extracted using our modified CHARMS framework revealed patterns in the identified studies. Appendix A.2 provides a detailed description of all extracted variables.

5. Results and discussion

5.1. Article selection

The database search yielded 1346 papers (Web of Science (n = 410; 30.5%), Scopus (n = 633; 47.0%), and MEDLINE (n = 303; 22.5%)), with 832 unique papers. Exclusions during the title/ abstract screening were mainly due to non-predictive studies (n = 250), lack of EHR use (n = 205), and techniques not relevant to our review research questions (n = 162). Title/ abstract screening identified 37 reviews/ background articles, where six papers were identified from forward citation screening for full-text screening [5,23–27].

Full-text screening identified 27 papers for data extraction. The PRISMA flowchart in Fig. 1 provides a summary of article selection; additional details are in Appendix A.3.

5.2. Risk of bias analysis

How do biases manifest within this literature, and what impact do they have on the reliability of study results?

Table 1
Risk of bias and applicability table formed from following the PROBAST guidelines.

	Risk of Bias (RoB)				Applicability			Overall	
Study	Participants	Predictors	Outcome	Analysis	Participants	Predictors	Outcome	RoB	Applicability
[5]	H	H	H	H	L	H	L	H	H
[6]	H	L	L	L	L	L	L	H	L
[29]	H	L	H	H	L	U	L	H	U
[9]	L	L	L	H	L	L	L	H	L
[8]	H	L	L	L	L	L	L	H	L
[30]	L	L	L	L	L	L	L	L	L
[15]	L	L	H	H	L	L	L	H	L
[16]	H	H	U	H	L	L	L	H	L
[31]	H	H	H	L	L	L	L	H	L
[32]	H	L	H	L	L	L	L	H	L
[14]	H	L	L	H	L	U	L	H	U
[33]	L	H	H	H	L	L	L	H	L
[34]	L	L	L	L	L	H	L	L	H
[35]	H	H	H	H	L	U	L	H	U
[17]	U	H	H	H	L	L	L	H	L
[36]	L	H	H	L	L	L	L	H	L
[37]	H	H	L	H	L	L	L	H	L
[38]	L	L	L	L	L	U	L	L	U
[39]	U	H	H	H	L	L	L	H	L
[40]	H	U	U	H	L	U	L	H	U
[41]	L	L	L	U	L	L	L	U	L
[7]	U	L	H	H	L	L	L	H	L
[42]	L	L	H	L	L	L	L	H	L
[43]	L	L	L	H	L	H	L	H	H
[44]	L	L	H	L	L	U	L	H	U
[45]	L	U	H	H	L	L	L	H	L
[46]	U	L	U	H	L	L	L	U	L

H- High risk, L - Low Risk, U - Unclear risk.

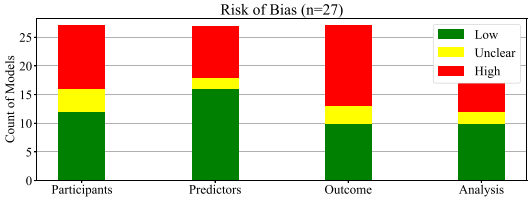

Fig. 2. Risk of bias of the papers included for data extraction.

Table 1 displays the RoB and applicability assessment at the domain and overall levels, with each row representing one study. Fig. 2 presents the breakdown of RoB levels.

5.2.1. Participant RoB

High RoB in the participant domain can be categorised into three groups. The first group have limited information about the target population, likely introducing bias based on available data [5,6,8,14,16,31,32,35,40]. This differs from papers with unclear RoB, where no information is given about data sources or the population [17,39], or where the control group is unclear [46]. The second group comprises papers with inclusion/exclusion criteria leading to inclusion bias [14,16,29,31,37]. These papers select or exclude patients based on specific characteristics relevant to the predictive context. The final group lacks control/comparison groups to determine prediction effectiveness [5,7].

5.2.2. Predictor RoB

Diagnostic codes in EHRs signify the presence of diseases and can serve as features for prediction or be the target labels. The International Classification of Diseases (* denotes version 9 or 10) (ICD*-CM), established by the World Health Organisation (WHO), is the global standard for diagnostic codes, facilitating systematic health outcome comparison across centres/institutions [47]. Its hierarchical structure provides varying levels of granularity, allowing the grouping of similar disease processes.

Incorrectly applying non-reproducible transformations or grouping medical events, International Classification of Diseases (ICD) codes, and medication codes poses a risk of misclassifying predictive features [16,17,31,33]. Utilising tools like OpenSAFELY helps find approved code lists for appropriate patient grouping [48].

False assumptions about the data context in EHRs can introduce inconsistencies and RoB. In acute settings, diagnoses are presumed, not confirmed, and should not be used as predictors [36]. Inappropriate imputation of missing data for ML algorithms can create unrealistic data given the clinical context and should not be used for prediction [37]. Additionally, self-reported lifestyle factors pose a significant risk of recall bias and should not be used as predictors [35].

The clinical utility of a predictive model depends on considering the timing of the prediction. Validity requires taking into account the availability of variables at the time of the prediction. Variables only available after the event time horizon should be avoided, as they render retrospective predictions clinically irrelevant [5,33,39].

Two papers lack sufficient information for RoB assessment in this domain. One lacks information about defined and assessed predictors [40], and the other lacks information about the timing of the index diagnosis [45].

These oversights indicate a lack of consideration for the clinical context in the design phase of model construction.

5.2.3. Outcome RoB

Approximately half of the papers showed high RoB in the outcome domain, falling into four groups: predictors shaping outcome definitions, flawed outcome assumptions, subjective outcome definitions, and poor methodology.

Consideration of the relationship between outcomes and predictors is essential. Specific predictors can unintentionally detect an outcome, such as using investigations or symptoms that are part of the diagnostic criteria (e.g., exacerbation of Chronic Obstructive–Pulmonary Disease (COPD) predicting COPD onset [17]), the presence of a treatment regimen predicting subsequent diagnosis [29], or tests specific to the cancer outcome in question [33]. Acknowledging the potential masking of

certain investigations when predicting disease onset should be explicit in the paper [16].

Constructing a composite outcome, like treatment failure, requires careful consideration. Some studies define treatment failure as a patient having the same diagnostic code within two weeks [15,36,42]. However, this assumes patients will return to the clinician within two weeks and that the initial diagnosis is correct. Such an approach is not a formal assessment of prescription effectiveness and should not be used as an outcome.

Despite ICD coding standardisation, choosing ICD codes or diagnoses as target variables can be ambiguous, risking observer bias [32]. Some ICD codes encompass a broad range of diagnoses (e.g., N39 – Other disorders of the urinary system) or lack agreed-upon standards (e.g., E86 – Volume depletion). Subjectivity also arises in applying specific diagnostic criteria, such as Alzheimer's disease and HF [7,44], or determining the cause of death [35].

Some papers have methodologically poor outcomes lacking predictive value in the clinical setting. Examples include papers with no prediction time horizon, rendering predictions irrelevant [5,7,31], or those with multi-outcome models predicting the following diagnosis or top k diagnoses [17,31].

5.2.4. Analysis RoB

Around half of the papers are high RoB in the analysis domain (n = 15; 56%); this domain had the highest number of high-risk scores. Reasons for high RoB are divided into papers at risk of being over-optimistic and those with inappropriate analysis of results.

Many methodologies lacked a specified number of participants in the outcome group(s) [7,15–17,29,33,37,43]. Without sample size calculations, it is unclear whether ML models have sufficient power to predict accurately, risking over-fitting and over-optimistic performance.

Data complexities like censoring and competing risks are not appropriately addressed or mentioned in the analysis [5,9,14,16,33,37,39,40,43,45,46]. Three papers lacked performance metrics or applied them inappropriately, such as using AUROC for highly unbalanced data [5,16,43]. One study treated alive patients differently from those who died at the end of the period [35].

5.2.5. Overall RoB

Only one paper, Golmaei et al. [30], has both an overall low RoB rating and low-risk applicability. Seven papers exhibit high or unclear RoB in a single domain, indicating that most literature in this area faces methodological issues across multiple domains leading to high RoB (n = 17; 63%).

Our RoB findings align with those of related works in clinical prediction modelling. Yang et al. conducted a systematic review of clinical prediction papers from 2009 to 2019, identifying 579 predictive models [49]. Navarro et al. performed a systematic review of 152 clinical prediction papers (models = 522) between 2018 and 2020, focusing on trends in methodological conduct reporting [50]. Navarro found that only a minority of papers performed external validation (12.5%), hyperparameter optimisation (28.9%), or provided calibration curves (5.4%).

This highlights a deficiency in the adherence of clinical prediction models to PROBAST guidelines. This could stem from a lack of awareness of PROBAST guidance or authors prioritising predictive model performance over applicability in a clinical setting. Addressing this gap is crucial for the effective use of these models in clinical practice.

5.2.6. Overall applicability

Three papers (11%) in our search are not applicable due to using population graphs instead of patient-level EHR representations [5,34,43]. Nineteen papers (70%) clearly use graphs to represent patient-level EHR data. The remaining six papers have unclear definitions of graph representations. Among them, four papers (15%) may use population-based representation [35,38,40,44], one paper represents graphs as single visits [14], and one paper is unclear about the graph representation [29].

5.3. Characteristics of included studies

Tables 11–15 in Appendix A.4 summarise the data extracted from the 27 papers.

5.3.1. Datasets and data sources

Eighteen datasets are used in the identified papers, broadly categorised as open source, dedicated research databases, and non-public/proprietary data. Table 2 provides a summary of the datasets used.

Three papers (11%) use simulated/synthetic data alongside a pre-existing dataset [7,14,35]. However, synthetic data inadequately captures complexities and relationships in EHRs, leading to poor representation of cohorts [51]. Consequently, we exclude these results, emphasising that these papers aim to demonstrate techniques rather than create clinical predictive models.

Healthcare delivery can broadly be divided into inpatient and outpatient/community care, each with distinct record structures and content reflecting the delivered care type. Outpatient records, likely sparse, offer better time coverage compared to sporadic but detailed inpatient records. When designing predictive models, these differences are crucial, as the clinical utility of predictions relies on potential levers for change in these settings. Additionally, accessing different parts of the EHR (inpatient vs outpatient) must be considered, given that typical clinical end-users lack universal access.

The distinction between primary and secondary care systems is not always evident, with many hospitals providing community services. While all papers in this study seemingly used EHRs from secondary care, details and prediction targets suggest community care records' use [6,7,9,15–17,33,35,36,39,42,44]. This raises concern as access to healthcare records varies, necessitating clarity on data requirements for model reproducibility.

Medical data, despite anonymisation, carries a risk of re-identifying subjects [52], countering the need for accessible datasets to verify and reproduce predictive models. The popularity of Medical Information Mart for Intensive Care (MIMIC), being freely accessible, reflects its status as a benchmark for verifying predictive model performance, despite its limitations of being critical care-focused.

The recent Goldacre Review supports the scale implementation of Trusted Research Environment (TRE), offering researchers a secure environment to access medical data for model development or verification [53]. This approach provides a secure yet accessible avenue for working with anonymised EHRs.

5.3.2. Model types

Sub-question (a): What graph approaches are researchers taking to predict these health outcomes?

For sub-question (a): Table 3 shows and describes the different ML and DL models used with graph representations of EHRs to make healthcare outcome and diagnosis predictions. Fig. 3 shows the model categories and within the extracted studies.

Recurrent Neural Network (RNN)-based models (Long-Short Term Memory (LSTM) and Gated Recurrent Unit (GRU)), excel with EHR data for handling sequential/temporal information. CNNs are employed for capturing spatial correlations. Combining RNNs and CNNs can aid in learning both temporal and spatial patterns.

DL is recommended for superior performance compared to other ML methods, given its capacity to capture intricate relationships. However, this complexity may not always identify uncertainties in data or model them, posing risks in healthcare settings where future predictions might suffer if the data distribution changes [11].

It is worth comparing graph and non-graph models to benchmark models, Fig. 4 shows the different models used for comparison.

Debate surrounds the trade-off between accuracy and computational cost in ML models. Gómez-Carmona et al. demonstrated an 80% reduction in computational effort with only a 3% decrease in accuracy [54].

Table 2
Summary of datasets used in the selected papers.

Papers	Dataset	Description	Source Country
Open Sources			
[37]	MIMIC-II	Clinical data related to patient admission to ICU, diagnoses ICD-9, and lab test results. Lab tests extracted every hour from admission.	USA
[8,30,31,38,41,44]	MIMIC-III	ICU short-term records, inpatient, discharge summaries	USA
[46]	MIMIC-IV	ICU short-term records, inpatient, discharge summaries	USA
[14,39,44,46]	eICU	ICD-9 and CPT procedure codes	USA
Research Databases			
[9,15,36,42]	Taiwanese National Health Insurance Research Database (NHIRD)	ICD9-CM	Taiwan
[33]	Foundation Medicine Inc and Mayo Clinic EHR	Oncology genetic reports, phenotypical data. Lab tests, diagnoses, medical and family history	USA
[32]	National registry data	–	UK
[35]	Taiwan National Death Registry	ICD-9 and ICD-10	Taiwan
[44]	NYU Langone Health	Long-term inpatient and outpatient EHRs	USA
Non-public/Proprietary Datasets			
[16,17,29]	Medical system from a city in North China	ICD-10 codes	China
[5,40]	Australian healthcare system	Admission information, diagnoses, procedures (ICD-10, DRG, AN-SNAP)	Australia
[39]	Paediatric EHR data from a tertiary care hospital in China	Symptoms, medical examination information, medication codes and diagnosis codes	China
[43]	IVF clinic from General Hospital in Seoul	Treatment records (age, stimulation type, use of Wallace, number of embryos transferred, symptoms)	South Korea
[6]	Private healthcare hospital admission data	ICD-10 codes and administrative data	–
[7,29,34,44]	Not provided	EHR Read codes	–
[45]	CardioNet	EHR Data from Seoul Asan Medical Center	South Korea

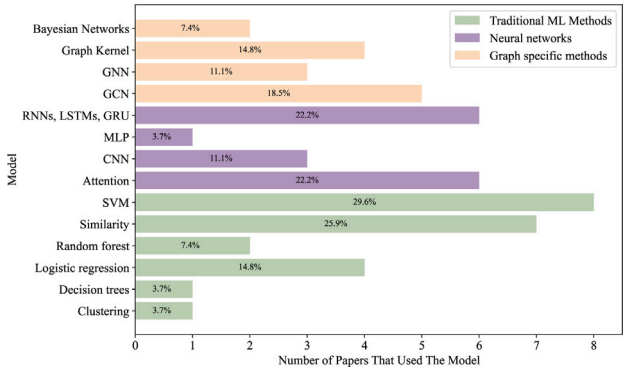

Fig. 3. The count of models and the percentage of papers within the selected 27 papers.

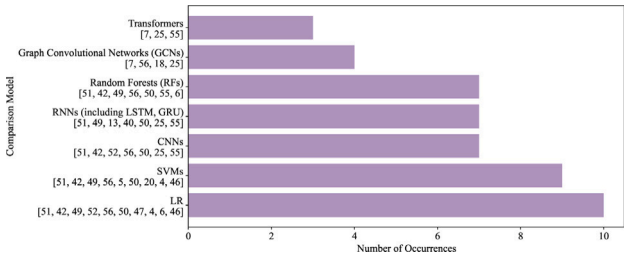

Fig. 4. Comparison/baseline model occurrence and their associated references (in square brackets).

None of the included papers supplied information on the time taken to train their models or the sample size. The exploration of resource intensity, model fitting, and prediction time for graph ML models represents a current gap in the literature. This information could be valuable in determining the feasibility of clinical implementation and optimising Pareto efficiency.

5.3.3. Graph representations

Sub-question (c): How are nodes and edges being utilised to perform these tasks?

We summarise sub-question (c) in Tables 4 and 5 which display node and edge types in the graphs, respectively. EHR read codes commonly served as node representations. Nodes were either homogeneous (41%) or heterogeneous (59%). Table 16 in Appendix A.5 details node and edge assignments in each model.

One of the proposed advantages of graph representation is the greater potential for explainability/interpretability. Graph representations allow data organisation in a non-linear path, providing explainable visualisation that would otherwise be complicated to infer. While minimising the 'black-box' effect of models is beneficial to provide explainability, especially in clinical settings, it is vital to consider that interpretability approaches may lead to artefacts from the learnt model rather than clinically explainable findings that should be attributed to the data [12].

There is a broad scope of interpretability, which to our knowledge, is without an encompassing definition. What is interpretable to an ML expert may not be to someone from a clinical background. Interpretability and transparency of models are essential to ensure that clinicians understand healthcare model choices to have confidence in decision-making rather than blind acceptance based on accuracy scores [11,55]. Several papers employed graph representations and ML for model interpretability. Some utilised graph kernels to enhance interpretability by calculating the similarity between patient graphs, assigning higher scores to patients with more matches with others [9,15,36,42]. Other models provided scores for features (nodes or edges), highlighting the most influential contributors to prediction decisions [15,31,33,40,44]. Additionally, graph visualisations were generated for clinicians to observe patient disease progression, aiding investigative direction [9,15,31,36,40]. Some suggested their graphs could aid in diagnosis and reveal causal relationships between EHR events [31].

Graph features exhibit various connection types to represent relationships. While typical Directed Acyclic Graphs (DAGs) have one-to-one links, more complex graphs, like hypergraphs, support one-to-many

Table 3
Descriptions of the different models used within the selected papers to make healthcare predictions. *DTs, LR, RF, SVM, GCNs, MLPs.*

Papers	Model Type	Description
Traditional ML Methods		
[5]	Clustering	Grouping similar data within a dataset using 2+ variables.
[6]	DTs	A single tree that makes predictions using previous answers. This forms a series of questions in a branched shape leading to the outcome.
[6,33,45,46]	LR	Linear classifier, which analyses the relationship between variables, using statistical analysis to predict binary outcomes. Note: the papers that use this method change the graph embeddings into vectorial representations.
[33,45]	RF	Multiple DTs trained via bagging techniques to optimise the predictive performance.
[5,6,29,32,34,37,40]	Similarity	Comparing 2+ samples to each other using distance or differences.
[7,9,15,17,33,40,42,46]	SVM	Used for both classification and regression, SVMs find the best hyperplane to divide the data into their groups.
Neural Networks		
[14,30,31,38,39,44]	Attention	Enabling attention to be paid to more valuable variables and reducing inefficiencies. It can also be used to show variables of importance and provide decision explainability.
[8,16,33]	CNN	Finds patterns in matrices (e.g., images, signal data) by applying filters and obtaining higher representations of the input data.
[33]	MLP	Neural network which is fully connected, the connections have varying weights which enforce or weakens connections to learn the patterns from input data.
[31,32,34,38,41,46]	RNNs, LSTM, GRU	Takes in sequential data and keeps it in memory by taking outputs from one step to the next. This means it has connections between time.
Graph Specific Methods		
[9,14,33,38,41]	GCNs	Like CNNs, GCNs learn using filters over data; however, GCNs can learn directly from nodes and their neighbouring nodes.
[30,35,44]	GNN	Neural networks can be used on graphs to analyse nodes, edges, relationships, and layouts to make predictions.
[9,15,36,42]	Graph Kernel	Convolution kernels on pairs of graphs, where the result from the convolution results in a new graph kernel.
[33,43]	Bayesian Network	Representation of conditional dependencies between variables using DAGs.

and many-to-many links. Analysing networks using methods like centrality or similarity measures reveals relationships between node and edge components. Modelling temporal EHR data is challenging due to irregularities, sparsity, heterogeneity, and model opacity [12], but graphs offer a way to address these challenges.

Another potential application of graph representation is utilising the temporal relationships of events within individual patient records that can examine sequences of events or event progression. By incorporating elapsed time as edge features or sequentially ordering events using directed graphs without specific time intervals, you can utilise the temporal dimension that can reveal valuable insights within the data, such time intervals between encounters may reveal patterns not obvious to clinicians [12].

While graph representations offer many advantages, we should be mindful that they can entail significant memory complexity and processing time, particularly in DL applications. This will be an important consideration when deploying models into a clinical environment, particularly during inference, where delayed predictions or insufficient compute resources are likely to lead to poor adoption.

5.3.4. Model performances

Sub-question (d): How do these graph approaches compare to each other? Sub-question (b): How do these approaches compare to other ML, Artificial Intelligence (AI), and statistical models?

Primary ML research methodology should offer sufficient performance metrics for independent evaluation. Across the papers, fourteen different metrics were provided. The most frequent metric was AUROC (70%), assessing model discrimination by comparing true positives to false positives. AUPRC followed as the second most used metric (56%), offering discriminative evaluation, particularly valuable in the presence of data imbalance. Accuracy (33%) provides a simple measure of correct predictions relative to all predictions. F1 score (26%) calculates the harmonic mean of precision and recall, preferred over accuracy in imbalanced data scenarios. Recall (26%) measures a model's ability to predict a positive outcome when present. Precision (22%) gives the positive predictive value. Specificity (7%) gauges a model's ability to predict a negative outcome when not present.

The following metrics appeared only once across all papers: Negative Predictive Value (NPV), coverage, true positives, true negatives, false positives, false negatives, and minimum precision and sensitivity. None of the papers included calibration curves or reported calibration, which helps detect overfitting by comparing predicted vs observed risk. Additionally, these papers lacked confidence intervals, a requirement for Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD) [56].

Of the surveyed articles, some provided binary classification alone (n = 17; 63%) , some multi-class (n = 2; 7%) and others gave risk/probability scores (n = 5; 19%). A few papers had multiple predictive tasks with both binary and multi-class classification (n = 3; 11%). Hospital readmission was the most popular prediction outcome (n = 9/43; 20.9%), followed by mortality (n = 9/43; 20.9%), and then treatment success (n = 4/43; 9.3%).

For sub-question (d) Table 6 displays models predicting mortality with AUROC or AUPRC scores. The highest AUROC score, 91.59%, was reported by Liu et al. [39]. Sun et al. [31] reported the best AUPRC score for mortality prediction at 81.34%. Regarding sub-question (d), Fig. 5 shows the AUROC and AUROC for the models which predict mortality. AUROC is a suitable metric to use if the dataset is balanced, however if it is not balanced AUPRC gives a better performance metric. The baseline score for AUPRC is typically determined by the prevalence of positive outcomes in the dataset. [31] have an unbalanced dataset, whilst [41] does not report dataset balance, which might explain the discrepancies between the AUROC and AUPRC scores.

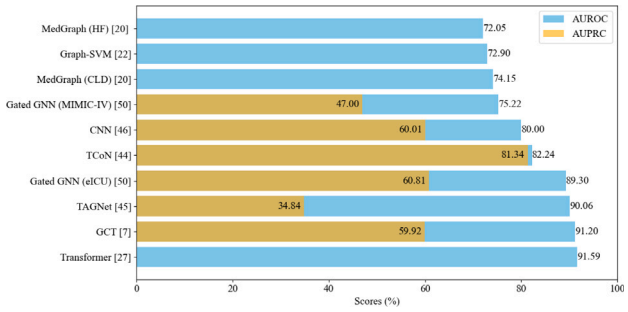

Fig. 5. AUROC and AUROC scores for the models predicting mortality.

For sub-question (d) Table 7 presents models predicting readmission with AUROC or AUPRC scores. The top-performing model for readmission prediction was reported by Golmaei et al. [30], achieving 85.8% ± 1.2 for AUROC and 84.7 ± 1.5 for AUPRC.

Tables 17–18 in Appendix A.6 display models predicting health outcomes, excluding mortality or readmission, with AUROC or AUPRC scores. Treatment success is a frequently predicted clinical outcome.

Due to dataset and validation method variations, we cannot perform quantified or statistical comparisons between the papers. However, among the three papers with low RoB predicting hospital readmission using the MIMIC-III dataset [30,38], GNN with Bi-directional Encoder Representation from Transformers (BERT) outperformed the GCN model with attention (AUROC + 3.3%, AUPRC + 21.5%). The higher-performing model underwent a more rigorous 5-fold Cross Validation (CV) validation, enhancing confidence in these results.

All included papers had main models outperforming or showing equivalent results to baseline predictive performance. However, despite this being relevant to sub-question (b)) this improvement regularly seen might be due to publication bias favouring papers with positive improvements. The high RoB suggests likely biased results and performance metrics, preventing statistical analysis or assessment of differences between primary and baseline models. A fair comparison would require identical predictive outcomes between models. As mortality was the most common outcome to be predicted we show the average AUROC difference between the comparison models and the main model from each of the 8 papers and 10 models within these in Figure 2 of Appendix A.6. From these comparisons we found that in most scenarios SVM and LSTM models alone have the largest difference in performance to the primary graph models.

Validation is sub-optimal with a single data split; CV or bootstrapping is preferred for calculating standard deviations and accommodating data variations during train/test splitting. External validation is lacking, representing a gap in implementing predictive models into clinical practice. Only four papers (15%) offered links to their GitHub repositories for model availability and reproducibility [31,34,44,46].

6. Limitations

Our search terms required specific search terms to capture the relevant literature. Due to the limited functionality of Google Scholar, we were unable to use this literature database and therefore may not have fully captured potential relevant literature.

While the initial aim was to examine how graph representations of EHRs are used, we only identified one applicable paper with low RoB, and so have focused on methodological bias. We highlight how current methods lead to high RoB with an aim to direct future research in this area to consider their assumptions to promote bias reduction; a must if any of these methods are to be clinically implemented.

The PROBAST tool, published in 2019, is the currently accepted guideline for assessing RoB within prediction model studies [21]. ML is categorised as predictive modelling and, in theory, falls into the scope

Table 4
Node allocation types in the graphs used in the selected papers. *UMLS*.

Papers	Node Allocation/Use	#
[5,7,8,14,16,17,29,33,34,36,39,45]	Diagnosis	12
[9,15,16,29,31,32,42,44]	EHR codes (e.g. ICD-10), but which type(s) are unclear e.g. diagnoses, demographic	8
[9,33–36,41,42]	Demographics (e.g. age, BMI, gender)	7
[7,17,29,33,36,45,46]	Medication	7
[8,14,17,33,45,46]	Laboratory investigations	6
[14,39,43,46]	Treatment	4
[8,33,45,46]	Patients	4
[35,39,45]	Physical examinations	3
[39,43]	Symptoms	2
[30,34]	Clinical note representation	2
[34,45]	Visits	2
[35]	Mental tests	1
[35]	Habits	1
[33]	Genetic data	1
[6]	Comorbidity occurrence count	1
[33]	Family history	1
[38]	Average values of word embeddings from: unique words from clinical free text or the linked UMLS	1
[37]	Discretised measurements of variables at a point in time	1
[40]	EHR Features	1
[41]	Heart rate, blood pressure and oxygen saturation	1
[41]	Eye-opening and verbal response	1
[45]	Smoking	1
[45]	Echocardiography	1

Table 5
Edge allocation types in the graphs used in the selected papers.

Papers	Edge Allocation	#
[9,15,17,34,36,42]	Time difference/elapsed between each node	6
[7,16,17,35]	Temporal proximity weighting	4
[5,6]	Number of times two diseases occurred simultaneously	2
[30,40]	The similarity between 2 nodes	2
[15,42]	Link to demographics (as the first node)	2
[5,37]	Sequential directionality/ ordering	2
[5,6]	Number of times two diseases occurred sequentially (one directly after another)	2
[41,44]	Fully connected initially and updated by attention	2
[14,33]	Association weighting between nodes	2
[45]	Relationship between patient and medical node e.g. edge exists between patient and smoke if the patient smokes	1
[29]	Weights higher if two medical events are more often and closer	1
[32]	Risk of disease	1
[34]	Different interactions, e.g. code to timestep	1
[38]	1) Intradocument interaction level. 2) Path lengths between entity nodes. 3) String similarities based on word overlap. 4) Cosine similarities	1
[43]	The conditional probability of a connection between 2 nodes	1
[39]	Medical relationship between nodes	1
[37]	Labelling of change of quantifiable variable (up, down or no change)	1
[31]	Linking of nodes/EHR codes happening on the same visit	1
[8]	Whether testing or diagnosis of a patient was undertaken	1
[46]	Events happening on the same time step are linked via edge and weighting is value from laboratory test, or infusion drug. If patient took a prescription the edge weight is 1 otherwise it is 0 to the prescription node	1

Table 6
Mortality prediction (binary) model performance. *GCT*.

Paper	Models used	Dataset	Classes/ Outcomes	AUROC (%)	AUPRC (%)	Validation Type
[8]	CNN	MIMIC-III	In-hospital mortality	80.00 ± 1	60.01 ± 1	10-fold CV
[31]	Co-occurrence-aware self-attention mechanism, Time-aware GRU	MIMIC-III	Mortality prediction	82.24	81.34	Train/ val/ test 0.75:0.1:0.15 5-fold CV
[14]	GCT	eICU	Mortality prediction	91.20 ± 0.48	59.92 ± 2.23	Train/val/test 8:1:1 split five times
[34]	Gaussian embedding, RNN	Not specified	Mortality prediction for: Heart failure Chronic liver disease	HF 72.05 Chronic liver disease 74.15	–	Train/val/test 80:15:5
[39]	Transformer	eICU	Mortality prediction	91.59	–	Train/val/test 8:1:1
[40]	L1-SVMs, Octagonal Shrinkage and Clustering Algorithm for Regression	Australian hospital	1 year mortality of cancer patients	72.90	–	Randomly divided into train and test sets 100 times
[41]	GRUs, attention, GCNs	MIMIC-III	Mortality in the next 24 h	90.06	34.84	Train/val/test 70:15:15
[44]	GNN, attention	MIMIC-III	Mortality prediction 24 h after admission	–	71.02	Train/val/test 8:1:1
[46]	Gated GNN	MIMIC-IV	Mortality caused by HF	75.22 ± 1.52	47.00 ± 2.13	5-fold CV
[46]	Gated GNN	eICU	Mortality caused by HF	89.30 ± 0.20	60.81 ± 0.76	5-fold CV

Table 7
Readmission prediction (binary) models with performance metrics. *ICU*.

Paper	Models used	Dataset	Classes/ Outcomes	AUROC (%)	AUPRC (%)	Validation Type
[30]	GNN, BERT	MIMIC-III	30 day hospital readmission	85.8 ± 1.2	84.7 ± 1.5	5-fold CV
[31]	Co-occurrence-aware self-attention mechanism, Time aware-GRU	MIMIC-III	Readmission	74.03	72.78	Train/val/test 0.75:0.1:0.15
[14]	GCT	eICU	Readmission during the same hospital stay	75.02 ± 1.14	52.44 ± 1.42	Train/val/test 8:1:1 split five times
[37]	Non-negative Matrix Factorisation	MIMIC-II	30 day ICU readmission risk	66.1	–	5-fold CV
[38]	GCN, attention, bi-directional LSTM	MIMIC-III	30 day ICU readmission risk	82.5	63.2	Train/val/test 8:1:1
[39]	Transformer	eICU	Readmission during a hospital stay	76.14	–	Train/val/test 8:1:1
[40]	L1-SVMs, Octogonal Shrinkage and Clustering Algorithm for Regression	Australian hospital	30 day hospital readmission Acute Myocardial Infarction	63.7	–	Randomly divided into train and test sets 100 times
[7]	SVM	Not specified	Risk of HF-related hospitalisation/ readmission	73	67	Random training and testing sets
[44]	GNN, attention	eICU	Readmission prediction at discharge	–	39.86	Train/val/test 8:1:1

of PROBAST. However, their complexity, in particular DL methods, means that the entire model cannot be fully presented. In contrast, statistical methods such as regression models can present coefficients. Different approaches and terminology within the field of ML mean that current PROBAST reporting may not fully critically appraise ML techniques, particularly throughout the analysis domain. There is ongoing work to address this potential pitfall, with new TRIPOD-AI and PROBAST-AI reporting guidelines anticipated, with a focus on ML methods [57]. Given that our RoB findings show that almost all papers were deemed high RoB within the analysis domain, it would

be expected that these papers would still have high RoB under new reporting guidelines.

7. Future directions

The papers in this review focus on demonstrating the utility of graph representation in improving predictive performance rather than clinical application. This is reflected by our RoB assessment which demonstrates that the literature in this area is making assumptions that would preclude its use in the clinical environment. Many of these papers fail to consider the clinical context of their prediction. These include using predictive variables that form part of the diagnostic criteria, the poor definition of clinical outcomes or using variables that occur only in the presence of the predicted outcome. Reporting conduct will only improve if authors and the bodies accepting these papers follow TRIPOD and PRISMA reporting guidelines [19,56].

We expect that such false assumptions would be addressed by having input from medical experts who understand the clinical context. Only 9 (33.3%) papers had clinical input in the paper, despite the papers having clinical predictive tasks. All the papers with a clinical author had a high RoB, suggesting they did not have the expertise to understand the RoB or did not have sufficient influence during the study design process. At this intersectional space of the application of Computer Science techniques within the domain of healthcare, better integration of medical expertise into predominately computer science teams may go some way in incorporating the clinical context and improving RoB.

None of these papers have formally explored how interpretable their analysis would be from the clinical end-user perspective and how this might change/affect clinical decision-making. Further research into formally defining the interpretability of predictive models and their effect on actionable change would be useful for graph representation and wider adoption of ML/AI within healthcare.

The ultimate goal of developing AI solutions within healthcare is to improve clinical outcomes. As with any ML modelling technique, prediction models must demonstrate robustness in other settings through external validation but must also be understood in the clinical context. The papers included in this review focus on the predictive aspect, which allows for earlier intervention and potentially better outcomes in some contexts. A larger question needs to be answered regarding the effect of improved predictions on clinical pathways and outcomes.

We have three key takeaways from our review. (1) Researchers should consider the clinical context carefully to ensure appropriate timing, code groupings, and a reasonable relationship between the outcome and predictors for clinical utility. (2) A lot of clinical research is not currently fit for clinical use due to researchers not following TRIPOD and PROBAST guidelines. The focus needs to shift from solely enhancing predictive modelling performance to improving the clinical utility of these models. (3) Graph representations have only been used for a limited number of purposes, there is further scope to expand graph models to other tasks. Graphs infrequently depict individual-level patient representation, and when employed, predictions are confined to just six outcomes (mortality, readmission, treatment success, sepsis, Cardiovascular disease (CVD), Alzheimer's), but we believe that graph usage could be extended to a wider range of health outcome prediction tasks such as utility or cancer recurrence etc.

The findings from reviewing these studies determine that methodological quality is poor, and a well-crafted health prediction paper has the following characteristics. It should be guided by the TRIPOD guidelines, ensuring transparency and reliability, minimising bias assessed through PROBAST. It begins by defining the research question and specifying the health outcome. The methodology should outline the predictors utilised within the model and the inclusion criteria, emphasising a representative sample. Rigorous internal validation by employing techniques like cross-fold validation and bootstrapping, is essential. The use of external validation gauges generalisability, and we acknowledge that this may extend beyond a single paper, necessitating follow-up studies by external teams. Model development relies on robust statistical methods, accounting for predictors and their interactions, and addresses missing data and biases. Transparent reporting, including calibration curves and confidence intervals, enhances result interpretability. Recent papers discuss and demonstrate some of the methods that should be used when creating models for healthcare applications [58–60].

8. Conclusion

Our review found 27 papers which used graph representation of EHRs for health outcome or prognosis prediction. A PROBAST analysis determined that only three papers had a low RoB. We present a narrative review of how EHR data can be represented as graphs by discussing characteristics of the methodologies, including model types, outcome prediction types, and model performances.

We found that researchers are mainly using 4 methods (GCNs, GNNs, Graph Kernels, and Bayesian Networks) to incorporate graphs into their healthcare prediction models. These graph approaches outperform baseline models that use non-graph ML, AI, and statistical methods. However this may potentially be due to publication bias. Diagnosis and EHR codes are most frequently being used for graph nodes, whilst edges are being used to represent time and simultaneous disease occurrence. Out of the 3 low RoB models, the GNN with BERT model had the best performance for hospital readmission prediction [30]. In the high RoB papers, the TCoN model (GNN with attention) had the best AUPRC performance.

Graph-based representations using EHRs, for individual health outcomes and diagnoses is an area ripe for exploration but require further knowledge building before we can see the results applied clinically. Graph representations appear to be useful in dealing with the sparsity of EHRs, by retaining structure and temporality. The simplicity of these graphs is also well suited for ML models for predicting health outcomes. Whilst mindful of publication bias, the technique of graph representation appears to improve predictive performance compared to baseline ML methods in multiple fields of medicine, suggesting the potential for universal application.

The high RoB suggests that authors do not use or are unaware of TRIPOD and PRISMA reporting guidelines. This may change with the publication of TRIPOD-AI and PROBAST-AI, which are specific to AI/ML methods [57]. These efforts are not insurmountable with a proper study design that incorporates clinical context, which will lead to suitable models within the clinical setting.

CRediT authorship contribution statement

Zoe Hancox: Writing – review & editing, Writing – original draft, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Allan Pang:** Writing – review & editing, Writing – original draft, Methodology, Investigation, Data curation. **Philip G. Conaghan:** Writing – review & editing, Supervision, Methodology, Conceptualization. **Sarah R. Kingsbury:** Writing – review & editing, Supervision, Methodology, Conceptualization. **Andrew Clegg:** Writing – review & editing, Supervision, Methodology, Conceptualization. **Samuel D. Relton:** Writing – review & editing, Supervision, Methodology, Conceptualization.

Funding

All the sources of funding for the work described in this publication are acknowledged below: Zoe Hancox is a PhD student supported through funding by the EPSRC (Grant No. EP/S024336/1). Allan Pang is a PhD student funded by the Ministry of Defence (United Kingdom).

Declaration of competing interest

No conflict of interest exists. We wish to confirm that there are no known conflicts of interest associated with this publication and there

has been no significant financial support for this work that could have influenced its outcome.

Acknowledgements

ZH is supported through funding by the EPSRC, United Kingdom (Grant No. EP/S024336/1). AP is funded by the Ministry of Defence (United Kingdom). PGC and SRK are funded in part by the National Institute for Health and Care Research (NIHR), United Kingdom through the Leeds Biomedical Research Centre. The views expressed are those of the authors and not necessarily those of the NHS, the NIHR or the Department of Health and Social Care.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.artmed.2024.102999.

References

[1] Puaschunder JM. The potential for artificial intelligence in healthcare. SSRN Electron J 2020;6(2):94–8. http://dx.doi.org/10.2139/ssrn.3525037.

[2] Liu X, Wang H, He T, Liao Y, Jian C. Recent advances in representation learning for electronic health records: A systematic review. J Phys Conf Ser 2022;2188(1):012007. http://dx.doi.org/10.1088/1742-6596/2188/1/012007.

[3] Hamilton WL. Graph representation learning. Synth Lect Artif Intell Mach Learn 2020;14(3):1–159. http://dx.doi.org/10.2200/S01045ED1V01Y202009AIM046, URL https://www.morganclaypool.com/doi/10.2200/S01045ED1V01Y202009AIM046.

[4] Scarselli F, Gori M, Tsoi AC, Hagenbuchner M, Monfardini G. The graph neural network model. IEEE Trans Neural Netw 2009;20(1):61–80. http://dx.doi.org/10.1109/TNN.2008.2005605.

[5] Khan A, Uddin S, Srinivasan U. Adapting graph theory and social network measures on healthcare data. In: Proceedings of the australasian computer science week multiconference. (February):New York, NY, USA: ACM; 2016, p. 1–7. http://dx.doi.org/10.1145/2843043.2843380, URL https://dl.acm.org/doi/10.1145/2843043.2843380.

[6] Khan A, Uddin S, Srinivasan U. Chronic disease prediction using administrative data and graph theory: The case of type 2 diabetes. Expert Syst Appl 2019;136:230–41.

[7] Liu C, Wang F, Hu J, Xiong H. Temporal phenotyping from longitudinal electronic health records: A graph based framework. In: Proceedings of the ACM SIGKDD international conference on knowledge discovery and data mining. 2015-Augus, 2015, p. 705–14. http://dx.doi.org/10.1145/2783258.2783352.

[8] Wanyan T, Honarvar H, Azad A, Ding Y, Glicksberg BS. Deep learning with heterogeneous graph embeddings for mortality prediction from electronic health records. Data Intell 2021;3(3):329–39. http://dx.doi.org/10.1162/dint_a_00097, arXiv:2012.14065.

[9] Yao HR, Chang DC, Frieder O, Huang W, Liang IC, Hung CF. Cross-global attention graph kernel network prediction of drug prescription. In: Proceedings of the 11th ACM international conference on bioinformatics, computational biology and health informatics, BCB 2020. Association for Computing Machinery, Inc; 2020, p. 1–10. http://dx.doi.org/10.1145/3388440.3412459, arXiv:2008.01868.

[10] Schrodt J, Dudchenko A, Knaup-Gregori P, Ganzinger M. Graph-representation of patient data: A systematic literature review. J Med Syst 2020;44(4):86.

[11] Xiao C, Choi E, Sun J. Opportunities and challenges in developing deep learning models using electronic health records data: A systematic review. J Am Med Inform Assoc 2018;25(10):1419–28. http://dx.doi.org/10.1093/jamia/ocy068.

[12] Xie F, Yuan H, Ning Y, Ong MEH, Feng M, Hsu W, et al. Deep learning for temporal data representation in electronic health records: A systematic review of challenges and methodologies. J Biomed Inform 2022;126(July 2021):103980. http://dx.doi.org/10.1016/j.jbi.2021.103980, arXiv:2107.09951.

[13] Si Y, Du J, Li Z, Jiang X, Miller T, Wang F, et al. Deep representation learning of patient data from electronic health records (EHR): A systematic review. J Biomed Inform 2021;115(October 2020):103671. http://dx.doi.org/10.1016/j.jbi.2020.103671, arXiv:2010.02809.

[14] Choi E, Xu Z, Li Y, Dusenberry MW, Flores G, Xue E, et al. Learning the graphical structure of electronic health records with graph convolutional transformer. In: AAAI 2020 - 34th AAAI Conference on Artificial Intelligence. 2020, p. 606–13. http://dx.doi.org/10.1609/aaai.v34i01.5400, arXiv:1906.04716.

[15] Yao HR, Chang DC, Frieder O, Huang W, Lee TS. Graph kernel prediction of drug prescription. In: 2019 IEEE EMBS international conference on biomedical and health informatics, BHI 2019 - proceedings. 2019, http://dx.doi.org/10.1109/BHI.2019.8834676.

[16] Zhang J, Gong J, Barnes L. HCNN: Heterogeneous convolutional neural networks for comorbid risk prediction with electronic health records. In: Proceedings - 2017 IEEE 2nd international conference on connected health: applications, systems and engineering technologies, CHASE 2017. 2017, p. 214–21.

[17] Zhang S, Liu L, Li H, Xiao Z, Cui L. Mtpgraph: A data-driven approach to predict medical risk based on temporal profile graph. In: 2016 IEEE trustcom/bigDataSE/ISPA, Vol. 1. IEEE; 2016, p. 1174–81. http://dx.doi.org/10.1109/TrustCom.2016.0191, URL http://ieeexplore.ieee.org/document/7847074/.

[18] Hossain ME, Khan A, Moni MA, Uddin S. Use of electronic health data for disease prediction: A comprehensive literature review. IEEE/ACM Trans Comput Biol Bioinform 2021;18(2):745–58. http://dx.doi.org/10.1109/TCBB.2019.2937862.

[19] Moher D, Shamseer L, Clarke M, Ghersi D, Liberati A, Petticrew M, et al. Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015 statement. Rev Esp Nutr Hum Diet 2016;20(2):148–60. http://dx.doi.org/10.1186/2046-4053-4-1.

[20] Ouzzani M, Hammady H, Fedorowicz Z, Elmagarmid A. Rayyan-a web and mobile app for systematic reviews. Syst Rev 2016;5(1):1–10. http://dx.doi.org/10.1186/s13643-016-0384-4, URL http://dx.doi.org/10.1186/s13643-016-0384-4.

[21] Wolff RF, Moons KG, Riley RD, Whiting PF, Westwood M, Collins GS, et al. PROBAST: A tool to assess the risk of bias and applicability of prediction model studies. Ann Intern Med 2019;170(1):51. http://dx.doi.org/10.7326/M18-1376, URL http://annals.org/article.aspx?doi=10.7326/M18-1376.

[22] Moons KG, de Groot JA, Bouwmeester W, Vergouwe Y, Mallett S, Altman DG, et al. Critical appraisal and data extraction for systematic reviews of prediction modelling studies: The CHARMS checklist. PLoS Med 2014;11(10). http://dx.doi.org/10.1371/journal.pmed.1001744, URL www.plosmedicine.org.

[23] McCormick T, Rudin C, Madigan D. A hierarchical model for association rule mining of sequential events: An approach to automated medical symptom prediction. 2011.

[24] Loscalzo J, Kohane I, Barabasi AL. Human disease classification in the postgenomic era: A complex systems approach to human pathobiology. Mol Syst Biol 2007;3(124). http://dx.doi.org/10.1038/msb4100163.

[25] Baglioni M, Pieroni S, Geraci F, Mariani F, Molinaro S, Pellegrini M, et al. A new framework for distilling higher quality information from health data via social network analysis. In: Proceedings - IEEE 13th international conference on data mining workshops, ICDMW 2013. 2013, p. 48–55.

[26] Folino F, Pizzuti C, Ventura M. A comorbidity network approach to predict disease risk. (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), Lecture Notes in Comput Sci 2010;6266 LNCS(i):102–9. http://dx.doi.org/10.1007/978-3-642-15020-3_10.

[27] Chang CD, Wang CC, Jiang BC. Using data mining techniques for multi-diseases prediction modeling of hypertension and hyperlipidemia by common risk factors. Expert Syst Appl 2011;38(5):5507–13. http://dx.doi.org/10.1016/j.eswa.2010.10.086, URL http://dx.doi.org/10.1016/j.eswa.2010.10.086.

[28] Haddaway NR, Page MJ, Pritchard CC, McGuinness LA. PRISMA2020: An r package and shiny app for producing PRISMA 2020-compliant flow diagrams, with interactivity for optimised digital transparency and open synthesis. Campbell Syst Rev 2022;18(2):e1230.

[29] Zhang S, Liu L, Li H, Cui L. Collaborative prediction model of disease risk by mining electronic health records. In: Lecture notes of the institute for computer sciences, social-informatics and telecommunications engineering, LNICST, Vol. 201. 2017, p. 71–82. http://dx.doi.org/10.1007/978-3-319-59288-6_7, URL http://link.springer.com/10.1007/978-3-319-59288-6{_}7.

[30] Golmaei SN, Luo X. DeepNote-GNN: Predicting hospital readmission using clinical notes and patient network. In: Proceedings of the 12th ACM conference on bioinformatics, computational biology, and health informatics, BCB 2021. Association for Computing Machinery, Inc; 2021, p. 1–9. http://dx.doi.org/10.1145/3459930.3469547.

[31] Sun C, Dui H, Li H. Interpretable time-aware and co-occurrence-aware network for medical prediction. BMC Med Inform Decis Mak 2021;21(1):1–12. http://dx.doi.org/10.1186/s12911-021-01662-z, URL https://doi.org/10.1186/s12911-021-01662-z.

[32] Qian Z, Alaa AM, Bellot A, Rashbass J, van der Schaar M. Learning dynamic and personalized comorbidity networks from event data using deep diffusion processes. In: Proceedings of the 23rd international conference on artificial intelligence and statistics 2020, Vol. 108. Palermo, Italy; 2020, p. 3295–305, arXiv:2001.02585.

[33] Zong N, Ngo V, Stone DJ, Wen A, Zhao Y, Yu Y, et al. Leveraging genetic reports and electronic health records for the prediction of primary cancers: Algorithm development and validation study. JMIR Med Inform 2021;9(5):1–18. http://dx.doi.org/10.2196/23586.

[34] Hettige B, Wang W, Li YF, Le S, Buntine W. MedGraph: Structural and temporal representation learning of electronic medical records. Frontiers Artificial Intelligence Appl 2020;325:1810–7. http://dx.doi.org/10.3233/FAIA200296, arXiv:1912.03703.

[35] Chen L, Li X, Sheng QZ, Peng WC, Bennett J, Hu HY, et al. Mining health examination records - a graph-based approach. IEEE Trans Knowl Data Eng 2016;28(9):2423–37. http://dx.doi.org/10.1109/TKDE.2016.2561278.

[36] Yao HR, Chang DC, Frieder O, Huang W, Lee TS. Multiple graph kernel fusion prediction of drug prescription. In: ACM-BCB 2019 - proceedings of the 10th ACM international conference on bioinformatics, computational biology and health informatics. 2019, p. 103–12. http://dx.doi.org/10.1145/3307339.3342134.

[37] Xue Y, Klabjan D, Luo Y. Predicting ICU readmission using grouped physiological and medication trends. Artif Intell Med 2019;95(3):27–37. http://dx.doi.org/10.1016/j.artmed.2018.08.004, URL https://linkinghub.elsevier.com/retrieve/pii/S0933365717306486.

[38] Lu Q, Nguyen TH, Dou D. Predicting patient readmission risk from medical text via knowledge graph enhanced multiview graph convolution. In: Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval. New York, NY, USA: ACM; 2021, p. 1990–4. http://dx.doi.org/10.1145/3404835.3463062, URL https://dl.acm.org/doi/10.1145/3404835.3463062.

[39] Liu X, Wang H, He T, Gong X. Research on intelligent diagnosis model of electronic medical record based on graph transformer. In: Proceedings - 2021 6th international conference on computational intelligence and applications, ICCIA 2021. IEEE; 2021, p. 73–8. http://dx.doi.org/10.1109/ICCIA52886.2021.00022.

[40] Kamkar I, Gupta S, Li C, Phung D, Venkatesh S. Stable clinical prediction using graph support vector machines. Proc Int Conf Pattern Recognit 2016;3332–7. http://dx.doi.org/10.1109/ICPR.2016.7900148.

[41] Wang S, Liu J. TAGNet: Temporal aware graph convolution network for clinical information extraction. In: 2020 IEEE international conference on bioinformatics and biomedicine. BIBM, IEEE; 2020, p. 2105–8. http://dx.doi.org/10.1109/BIBM49941.2020.9313530, URL https://ieeexplore.ieee.org/document/9313530/.

[42] Chang D-C, Frieder O, Hung C-F, Yao H-R. The analysis from nonlinear distance metric to kernel-based prescription prediction system. J Nonlinear Var Anal 2021;5(2):179–99. http://dx.doi.org/10.23952/jnva.5.2021.2.01, URL http://jnva.biemdas.com/archives/1308.

[43] Kim IC, Jung YG. Using Bayesian networks to analyze medical data. (Subseries of Lecture Notes in Computer Science), Lecture Notes in Artificial Intelligence 2003;2734:317–27. http://dx.doi.org/10.1007/3-540-45065-3_28.

[44] Zhu W, Razavian N. Variationally regularized graph-based representation learning for electronic health records. ACM CHIL 2021 - proceedings of the 2021 ACM conference on health, inference, and learning, Vol. 1. Association for Computing Machinery; 2021, p. 1–13. http://dx.doi.org/10.1145/3450439.3451855, arXiv:1912.03761.

[45] Cho HN, Ahn I, Gwon H, Kang HJ, Kim Y, Seo H, et al. Heterogeneous graph construction and hinsage learning from electronic medical records. Sci Rep 2022;12(1):1–9. http://dx.doi.org/10.1038/s41598-022-25693-2, URL https://doi.org/10.1038/s41598-022-25693-2.

[46] Xu Y, Ying H, Qian S, Zhuang F, Zhang X, Wang D, et al. Time-aware context-gated graph attention network for clinical risk prediction. IEEE Trans Knowl Data Eng 2022;14(8). http://dx.doi.org/10.1109/TKDE.2022.3181780.

[47] Harrison JE, Weber S, Jakob R, Chute CG. ICD-11: An international classification of diseases for the twenty-first century. BMC Med Inform Decis Mak 2021;21(6):1–10.

[48] OpenSAFELY. About opensafely. 2022, URL https://www.opensafely.org/about/.

[49] Yang C, Kors JA, Ioannou S, John LH, Markus AF, Rekkas A, et al. Trends in the conduct and reporting of clinical prediction model development and validation: A systematic review. J Am Med Inform Assoc 2022;29(5):983–9. http://dx.doi.org/10.1093/jamia/ocac002.

[50] Andaur Navarro CL, Damen JA, van Smeden M, Takada T, Nijman SW, Dhiman P, et al. Systematic review identifies the design and methodological conduct of studies on machine learning-based prediction models. J Clin Epidemiol 2023;154:8–22. http://dx.doi.org/10.1016/j.jclinepi.2022.11.015, URL https://doi.org/10.1016/j.jclinepi.2022.11.015.

[51] Murray RE, Ryan PB, Reisinger SJ. Design and validation of a data simulation model for longitudinal healthcare data. AMIA Annu Symp Proc Arch 2011;1176–85.

[52] Gille F, Brall C. Limits of data anonymity: Lack of public awareness risks trust in health system activities. Life Sci Soc Policy 2021;17(1):1–8.

[53] Goldacre B, Morley J. Better, Broader, Safer: Using Health Data for Research and Analysis. Technical Report, Department of Health and Social Care; 2022, URL https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment{_}data/file/1067053/goldacre-review-using-health-data-for-research-and-analysis.pdf?utm{_}campaign=846512{_}PRESSRELEASEGoldacrereview&utm{_}medium=email&utm{_}source=NHSConfe.

[54] Gómez-Carmona O, Casado-Mansilla D, Kraemer FA, López-de Ipiña D, García-Zubia J. Exploring the computational cost of machine learning at the edge for human-centric Internet of Things. Future Gener Comput Syst 2020;112:670–83. http://dx.doi.org/10.1016/j.future.2020.06.013, URL https://linkinghub.elsevier.com/retrieve/pii/S0167739X20304106.

[55] Evans RP, Bryant LD, Russell G, Absolom K. Trust and acceptability of data-driven clinical recommendations in everyday practice: A scoping review. Int J Med Inf 2024;183:105342. http://dx.doi.org/10.1016/j.ijmedinf.2024.105342, URL https://www.sciencedirect.com/science/article/pii/S1386505624000054.

[56] Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): The TRIPOD statement. BMC Med 2015;13(1):1–10. http://dx.doi.org/10.1186/s12916-014-0241-z.

[57] Collins GS, Dhiman P, Navarro CLA, Ma J, Hooft L, Reitsma JB, et al. Protocol: Protocol for development of a reporting guideline (TRIPOD-AI) and risk of bias tool (PROBAST-AI) for diagnostic and prognostic prediction model studies based on artificial intelligence. BMJ Open 2021;11(7).

[58] Collins GS, Dhiman P, Ma J, Schlussel MM, Archer L, Calster BV, et al. Evaluation of clinical prediction models (part 1): From development to external validation. BMJ 2024;384. http://dx.doi.org/10.1136/bmj-2023-074819, URL https://www.bmj.com/content/384/bmj-2023-074819. arXiv:https://www.bmj.com/content/384/bmj-2023-074819.full.pdf.

[59] Riley RD, Archer L, Snell KIE, Ensor J, Dhiman P, Martin GP, et al. Evaluation of clinical prediction models (part 2): How to undertake an external validation study. BMJ 2024;384. http://dx.doi.org/10.1136/bmj-2023-074820, URL https://www.bmj.com/content/384/bmj-2023-074820. arXiv:https://www.bmj.com/content/384/bmj-2023-074820.full.pdf.

[60] Riley RD, Snell KIE, Archer L, Ensor J, Debray TPA, van Calster B, et al. Evaluation of clinical prediction models (part 3): Calculating the sample size required for an external validation study. BMJ 2024;384. http://dx.doi.org/10.1136/bmj-2023-074821, URL https://www.bmj.com/content/384/bmj-2023-074821. arXiv:https://www.bmj.com/content/384/bmj-2023-074821.full.pdf.