Journal of the American Medical Informatics Association, 2023, 1–11
https://doi.org/10.1093/jamia/ocad168
Review

Review

Prediction models using artificial intelligence and longitudinal data from electronic health records: a systematic methodological review

Lucía A. Carrasco-Ribelles, MSc[1,2,3], José Llanes-Jurado, MSc[4], Carlos Gallego-Moll, MSc[1,3], Margarita Cabrera-Bean, PhD[2], Mònica Monteagudo-Zaragoza, PhD[1], Concepción Violán, MD, PhD[3,5,6,7,*], Edurne Zabaleta-del-Olmo, PhD[1,8,9]

[1]Fundació Institut Universitari per a la recerca a l'Atenció Primària de Salut Jordi Gol I Gurina (IDIAPJGol), Barcelona, 08007, Spain, [2]Department of Signal Theory and Communications, Universitat Politècnica de Catalunya (UPC), Barcelona, 08034, Spain, [3]Unitat de Suport a la Recerca Metropolitana Nord, Fundació Institut Universitari per a la recerca a l'Atenció Primària de Salut Jordi Gol I Gurina (IDIAPJGol), Mataró, 08303, Spain, [4]Instituto de Investigación e Innovación en Bioingeniería (i3B), Universitat Politècnica de València (UPV), València, 46022, Spain, [5]Direcció d'Atenció Primària Metropolitana Nord, Institut Català de Salut, Badalona, 08915, Spain, [6]Fundació Institut d'Investigació en ciències de la salut Germans Trias i Pujol (IGTP), Badalona, 08916, Spain, [7]Fundació UAB, Universitat Autònoma de Barcelona, Cerdanyola del Vallès, 08193, Spain, [8]Gerència Territorial de Barcelona, Institut Català de la Salut, Carrer de Balmes 22, Barcelona, 08007, Spain, [9]Nursing Department, Faculty of Nursing, Universitat de Girona, Girona, 17003, Spain

*Corresponding author: Concepción Violán, MD, PhD, Unitat de Suport a la Recerca Metropolitana Nord, Fundació Institut Universitari per a la Recerca a l'Atenció Primària de Salut Jordi Gol I Gurina (IDIAPJGOL), Carrer Mare de Déu de Guadalupe 2, Mataró, 08303, Spain (cviolanf.mn.ics@gencat.cat)

Abstract

Objective: To describe and appraise the use of artificial intelligence (AI) techniques that can cope with longitudinal data from electronic health records (EHRs) to predict health-related outcomes.

Methods: This review included studies in any language that: EHR was at least one of the data sources, collected longitudinal data, used an AI technique capable of handling longitudinal data, and predicted any health-related outcomes. We searched MEDLINE, Scopus, Web of Science, and IEEE Xplorer from inception to January 3, 2022. Information on the dataset, prediction task, data preprocessing, feature selection, method, validation, performance, and implementation was extracted and summarized using descriptive statistics. Risk of bias and completeness of reporting were assessed using a short form of PROBAST and TRIPOD, respectively.

Results: Eighty-one studies were included. Follow-up time and number of registers per patient varied greatly, and most predicted disease development or next event based on diagnoses and drug treatments. Architectures generally were based on Recurrent Neural Networks-like layers, though in recent years combining different layers or transformers has become more popular. About half of the included studies performed hyperparameter tuning and used attention mechanisms. Most performed a single train-test partition and could not correctly assess the variability of the model's performance. Reporting quality was poor, and a third of the studies were at high risk of bias.

Conclusions: AI models are increasingly using longitudinal data. However, the heterogeneity in reporting methodology and results, and the lack of public EHR datasets and code sharing, complicate the possibility of replication.

Registration: PROSPERO database (CRD42022331388).

Key words: artificial intelligence; deep learning; electronic health records; longitudinal data; prediction; systematic review.

Introduction

Artificial intelligence (AI) is increasingly used in healthcare. Applications range from identifying the presence or predicting the development of a condition (eg, from medical imaging or electronic health records [EHR]) to treatment (eg, AI-guided robots that perform surgery), drug production (ie, bioinformatics), and training.[1] The progression of the diseases (eg, decompensation, mortality, etc.), which the next event (eg, prescription or diagnostic) will be, or measures for quality care (eg, hospital readmission, length of stay, etc.) can also be predicted. These techniques can be used in a variety of disciplines, such as radiology, oncology, or surgery.[2] Some AI-based models detect or predict health- or healthcare-related outcomes based on cross-sectional data, with satisfactory performance.[3,4] However, few use longitudinal data, which have numerous advantages in terms of quantity and quality and can show temporal changes in patients' conditions.[5,6] Temporal data collection can allow the study of causality between events and dynamics to analyze how people develop diseases or how some diseases develop into others. Understanding patients' evolution can be crucial for a correct prognosis of certain long-term conditions and could improve the performance of the AI models.[7]

The development of EHR systems has increased the number of studies using longitudinal data, but knowledge about AI

Received: May 5, 2023; **Revised:** August 2, 2023; **Editorial Decision:** August 9, 2023; **Accepted:** August 11, 2023

techniques using these data is still lacking. EHRs are a diverse source of data, as each patient's record can be different in terms of number, type, and frequency of registers. The healthcare setting in which the data was collected might influence these parameters. For instance, primary care records may reflect a patient's history over several years but with few annual visits usually of a single type. Conversely, ICU and hospitalization records might have short-term high-frequency registers of many different types, such as procedures, laboratory results, and diagnoses, but may lack information on the patient's life outside the hospital. Differences between subjects in the amount of data available can be handled in different ways (ie, adding time between events, aggregating events, or limiting the usable data in a fixed time window[8]), which sometimes makes the implementation technically difficult. Therefore, data from different settings are likely to need different methodological approaches and be more powerful to predict different outcomes, ie, ICU data might predict better short-term decompensation, while the life-long data from primary care might predict better quality care outcomes or long-term outcomes such as all-cause mortality.[9,10]

Different approaches have been used to take into account the temporal dimension of the data.[8] Even though some machine learning (ML) models can be assembled to handle longitudinality,[11,12] recurrent neural networks (RNN), and particularly long short-term memory (LSTM)[13] and gated recurrent units (GRU)[14] were designed to handle temporal sequenced data. These types of networks can consider the changes between sequential registers thanks to the incorporation of forget gates, and successfully address the vanishing gradient problem even in long sequences. Lately, RNN-based architectures have been enhanced by incorporating new types of layers, such as convolutional neural networks (CNN) or graph neural networks (GNN).[15,16] In addition, self-attention mechanisms can also be included in these architectures to improve both performance and interpretability by calculating attention weights that highlight the parts of the sequence that contribute most to the prediction.[17–19] These mechanisms have led to the development of pretrained systems like BERT that ease model training, which can also be applied to healthcare.[20]

Some systematic reviews have addressed the application of AI in healthcare,[3,4,21] but without a specific focus on longitudinal data, so the techniques reported may not be apt for the temporal data. Identifying the most common and successful AI techniques for longitudinal studies is key to improving the more common cross-sectional prediction AI models. This systematic methodological review aims to describe and appraise which AI techniques can cope with longitudinal data from EHRs to predict health- or healthcare-related outcomes. Specifically, we report the most common data specifications, the techniques used, how they are trained, what results they achieve, and the quality of the reporting and the methods used. By doing so, we sought to provide an overview of the state of the art in training AI models with longitudinal data.

Methods

This review is reported according to PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines.[22] The review protocol was prospectively registered in the PROSPERO database (CRD42022331388).

Review questions

Regarding AI techniques able to use longitudinal data from EHRs to make predictions, the following questions arose: (1) which data specifications are most common?, (2) which techniques are used?, (3) how are they trained?, (4) what results do they achieve?, and (5) what is the quality of the reporting and the risk of bias in the included studies?

Eligibility criteria

We followed the SDMO (Studies, Data, Methods, Outcomes) approach to define the eligibility criteria[23] (Supplementary Appendix A). We included studies published in any language that: (1) at least one data source was EHRs, (2) data collection implied having longitudinal data (ie, at least 2 registers per individual at different time points), (3) used an AI technique capable of handling longitudinal data (eg, RNN, Hidden Markov Models); and (4) made a prediction about a health-related outcome (eg, hospital admission, death, nursing home admission).

We excluded short reports of less than 2 pages because the reporting in this type of publication is highly variable in terms of reliability, accuracy, and level of detail. They often provide little information on study design and risk of bias, complicating their appraisal.[24–26] We also excluded records that originally had longitudinal data but transformed them in such a way (ie, averaging all the laboratory results over time) that the longitudinality was lost.

Information sources and search strategy

We systematically searched MEDLINE (PubMed), Web of Science, Scopus, and IEEE Xplorer from database inception to January 3, 2022, with no language restrictions. The search strategies run in each database can be found in Supplementary Appendix A (Table A1). We handsearched the reference lists of included studies and contacted experts in the area to identify other possible eligible studies.

Selection process

The search results were uploaded to Rayyan,[27] a web-based software, and deduplicated. Pairs of reviewers (of L.A.C.-R., CV, E.Z.O., M.M.-Z.) independently screened titles and abstracts against the eligibility criteria. We retrieved the full text of all studies that potentially met the eligibility criteria, and pairs of review authors (L.A.C.-R., C.V., E.Z.-O.) assessed them for inclusion, recording the reasons for exclusion. We resolved any disagreements through discussion, involving a third reviewer when needed.

Data collection process and data items

One reviewer (L.A.C.-R.) extracted data from included studies using a standardized form, and another reviewer (J.L.-J. or C.G.-M.) checked it. All reviewers had piloted the data extraction form on 5 articles to ensure consistency. Disagreements were resolved through consensus.

Collected data included information on the biggest dataset, prediction task, preprocessing of input data, feature selection, method, validation, performance metrics, and implementation. The definition of each variable is presented in Supplementary Appendix B (Table B1). In single studies that reported several models, the performance metrics of the best model were extracted. In addition, the transparent reporting of a multivariable prediction model for Individual Prognosis

Or Diagnosis (TRIPOD) statement was used to assess the transparent reporting of the studies.[28]

Risk of bias assessment

Pairs of reviewers (of L.A.C.-R., M.C.-B., E.Z.-O., C.V.) independently assessed risk of bias using a short-form of the Prediction model Risk Of Bias ASsessment Tool (PROBAST) statement[29] which has a sensitivity of 98% and specificity of 100% compared to the long form. We used the short form because it did not assess the risk of bias in the data, as EHR data are assumed to be at high risk of bias and all the included studies in this review used EHRs. In addition, risk of bias was assessed only in the articles that included the items necessary for its evaluation according to TRIPOD (see Supplementary Appendix C). These items referred to the outcome and analysis sections, and the overall risk of bias was assigned according to these sections. The extensions for TRIPOD and PROBAST for AI models were under development and unavailable at the time of this review.[30]

Descriptive statistics (ie, percentages, or medians and interquartile range, as appropriate) were calculated from the data extraction table. This data synthesis was performed using R (version 4.1.2).

Results

Of 391 records, 228 reports underwent full-text review, as 12 could not be retrieved (Figure 1). Among the 81 studies that met the eligibility criteria, 78 (96.3%) were published from 2017 to 2021. The raw parameters of included studies (see Supplementary Appendix D), together with a dashboard to explore the results in more detail, is available in a Shiny app (https://lacarrascoribelles-idiapjgol.shinyapps.io/SR_AILongitudinalModels/). Supplementary Appendix E lists the studies not retrieved, those excluded, and the reasons for exclusion.

Dataset information and prediction task

The included studies followed patients for 0.8-24 years, had sample sizes of 398-29 163 297 individuals, and used most of them EHR data from the United States. Simple sample demographic descriptors such as age, sex, or socioeconomic information were reported in less than 30% of the studies (Table 1). Although the sample was most commonly drawn from the general patient population, some studies defined an age limit (eg, ≥18, ≥65 years) or included only people with particular diseases (eg, heart failure, diabetes). None of these diseases were considered when defining the population in more than 3 studies. The most common settings were ICUs and hospitalization (Table 2).

In general, the studies simultaneously considered information from at least 3 of the following: diagnoses, billed drugs, demographics, medical procedures, and clinical measures. Diagnoses and drugs were the most common types of data used. Others, such as claims data, genes, or free text were considered in 17/81 (21%). Data were, in most cases, coded either as categorical or both quantitative and categorical variables. While most datasets did not have a particular individual recording frequency, some reported having one register per person every hour, day, or 2 years (Table 2). The mean number of registers per person varied between 1.2 and 82.5 (mean = 12.87, standard deviation [SD]=24.08), while 45/81 (55.6%) did not report it this variable.

Prediction tasks were numerous ($n = 67$) and heterogeneous. The most common were mortality, next diagnosis, and heart failure (Table 2). Most studies considered only one prediction task (58/81, 71.6%), while others used the same

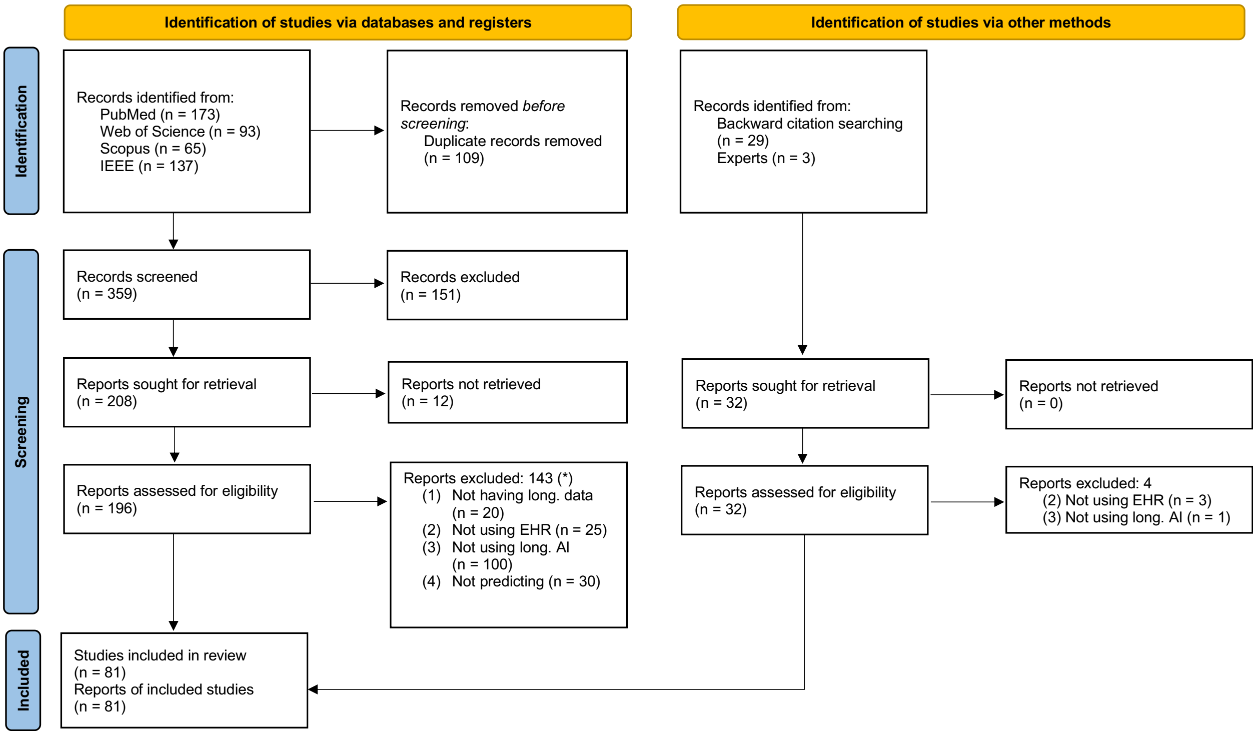
Identification of studies via databases and registers
Identification of studies via other methods
Identification
Records identified from:
PubMed (n = 173)
Web of Science (n = 93)
Scopus (n = 65)
IEEE (n = 137)
Records removed before screening:
Duplicate records removed (n = 109)
Records identified from:
Backward citation searching (n = 29)
Experts (n = 3)
Screening
Records screened (n = 359)
Records excluded (n = 151)
Reports sought for retrieval (n = 208)
Reports not retrieved (n = 12)
Reports sought for retrieval (n = 32)
Reports not retrieved (n = 0)
Reports assessed for eligibility (n = 196)
Reports excluded: 143 (*)
(1) Not having long. data (n = 20)
(2) Not using EHR (n = 25)
(3) Not using long. AI (n = 100)
(4) Not predicting (n = 30)
Reports assessed for eligibility (n = 32)
Reports excluded: 4
(2) Not using EHR (n = 3)
(3) Not using long. AI (n = 1)
Included
Studies included in review (n = 81)
Reports of included studies (n = 81)

Figure 1. PRISMA flow chart.[22] *The same study can be excluded for several reasons, so the sum of excluded studies for each reason does not add up to the total number of excluded studies.

Table 1. Description of the parameters related to the dataset information.

Parameter	Description
Country of data source	
United States	50 (61.7)
China	13 (16.0)
Others	15 (18.5)
Not reported	3 (3.7)
Follow-up time, in years	[0.8, 24], 8.9 (5.6), 10 [4–12], (70.4)
Sample size	[398, 29 163 297], 585 981 [3 611 801], 32 221 [7845–105 805], (18.5)
Demographic descriptors	
Median age	[29, 74.6], 55.6 (13.4), 56.1 [46.2–65.4], (23.5)
Sex (female)	[0, 100], 51 (20.6), 49.8 [43–61.4], (29.6)
Socioeconomic information reported	6 (7.4)
Age limit definition	19 (23.5)
Main condition definition	27 (8.6)
Information included	
Demographics	39 (48.2)
Diagnoses	60 (74.1)
Laboratory results	38 (46.9)
Prescribed/billed drugs	43 (53.1)
Medical procedures	31 (38.3)
Clinical measures	31 (38.3)
Information included simultaneously (*n*)	[1, 6], 3 (1.3), 3 [2–4], (98.8)
Others (eg, free text, genes...)	17 (21)
Types of variables	
All quantitative	7 (8.6)
All categorical	39 (48.1)
Both	35 (43.2)

Categorical parameters are described as *N* (%), while quantitative parameters as [min, max], mean (SD), median [Q1-Q3], (% studies reported). Total *N* = 81. If more than one dataset was used to train the models, only the parameters of the biggest dataset were collected.

architecture to predict up to 6 outcomes. The prediction window varied from 0 h to 10 years. Most studies only considered one prediction window (62/81, 76.54%), while some considered up to 5. Some studies (14/81, 17.3%) changed the distribution of the target outcome, either by performing a case-control selection, stratified random sampling, or under/oversampling.

Techniques

Almost a third of the studies (26/81, 32.1%) reported missing values in their dataset (Table 3), but few reported using compatible architectures. Most imputed them with different techniques (eg, last value carried forward, mean or median imputation, zero imputation). Only 2 studies reported indicating these imputations to the model through additional variables or masking.

All but 3 studies reported having preprocessed the data, via: one-hot encoding of categorical variables, embedding of some kind, aggregating the information available in time windows, removing the events (ie, diagnoses, procedures) that were less common, performing any kind of normalization on the quantitative variables, and categorizing the quantitative variables. Studies simultaneously performed up to 5 of these preprocessing procedures. In addition, 6/81 (7.4%) reported performing a feature selection, either according to statistical criteria (ie, Lasso, Mann-Whitney, Chi^2), expert's opinion, or missingness.

All studies but 2 that were based on ML methods had deep learning (DL) architectures (79/81, 97.5%) (Table 2). Most DL models were based on RNN (eg, LSTM/BiLSTM, GRU/BiGRU) architectures. Others were based either on BERT-based architectures or combinations of RNN-like layers with other architectures such as CNNs, GNNs, or autoencoders. Architectures were reported as having 1 up to 10 layers, but most were based on 1-3, while 32/81 (39.5%) studies did not report it. Table 3 shows that some kind of attention mechanism was included in 45/81 (55.6%) studies. Moreover, the architecture of 27/81 (33.3%) models could consider static variables, like sex. Except for the 5/81 (6.2%) studies that reported having the same number of registers per individual due to the study design, the rest applied some technique to make their model work varying numbers of registers per person. Most applied some preprocessing technique (31/81, 38.3%), such as aggregation in time windows. Zero-padding was used in 22/81 (27.2%) studies (masking was reported in 8/22, 36.4%), while 19/81 (23.5%) studies did not report how they managed this.

Hyperparameter tuning was reported in 36/81 (44.4%) studies. The most common techniques were fine-tuning, grid search, and Bayesian optimization. One to eight hyperparameters were tuned, most commonly the number of neurons per layer, the learning rate, the dropout rate, and the number of layers. Seven of the 36 (19.4%) studies did not report which hyperparameters tuned. Regarding regularization mechanisms (eg, dropout, L1 or L2 regularization), 52/81 (64.2%) studies applied up to 5 of them simultaneously, while 29/52 studies (54.7%) considered only one, and 19/52 (35.8%) two (Table 4). The most common optimizers were Adam, Stochastic Gradient Descent, and Adadelta.

Most studies (79/81, 97.5%) reported doing an internal validation of their models (Table 4), generally through either a random split of the dataset or a replication-based technique (eg, cross-validation, bootstrap) to estimate the error dispersion of the models. The size of the training set ranged from 14% to 90% (mean 73.8%, SD 10.7%), of the validation set from 4.4% to 50% (mean 14.5%, SD 7.5%), and of the test set from 0% to 80% (mean 15%, SD 11%). External validation was reported in 3/81 (3.7%) studies: 2 using the MIMIC-III dataset, and one a private one. In 37/81 (45.7%) studies, performance of the developed models was compared to up to 8 other state-of-the-art methods on the same dataset. The most common were RETAIN, Dipole, and SAnD. In addition, 70/81 (86.42%) studies compared the performance of their models to up to 12 simpler methods, which usually could not benefit from the temporal dimension of their data, such as logistic regression or random forests.

Regarding the implementation of the models, 22/81 (27.2%) studies reported using Tensorflow; 15/81 (18.5%) PyTorch; and 10/81 (12.4%) Keras. The framework used was not reported in 21/81 (25.9%) studies. Thirty-one studies (38.3%) made some of the code related to their model available in Git-like platforms.

Data available and techniques used according to data context

At least half the studies in each setting had as many registers per patient as patient contacts with the health system

Table 2. Description of the registers collection, the prediction task, and the architecture by data context.

		Hospital care ($N=52$)					
Parameter	Primary care ($N=7$)	Consultation ($N=4$)	Hospitalization ($N=20$)	ICU ($N=28$)	Multiple healthcare settings ($N=11$)	Not reported ($N=11$)	Total ($N=81$)
Mean number registers per person							
0-10	0 (0.0)	0 (0.0)	5 (25.0)	8 (28.6)	2 (18.2)	2 (18.2)	17 (21.0)
11-50	0 (0.0)	2 (50.0)	2 (10.0)	2 (7.1)	1 (9.1)	4 (36.4)	11 (13.6)
51-100	3 (42.9)	0 (0.0)	3 (15.0)	2 (7.14)	0 (0.0)	0 (0.0)	8 (9.9)
Not reported	4 (57.1)	2 (50.0)	10 (50.0)	16 (57.1)	8 (72.7)	5 (45.5)	45 (55.6)
Frequency of registers ($N=84$)	($N=9$)	($N=4$)	($N=20$)	($N=29$)	($N=11$)	($N=11$)	($N=84$)
Hourly (every 0.5, 1, 2, … h)	0 (0.0)	0 (0.0)	2 (10.0)	8 (27.6)	2 (18.2)	0 (0.0)	12 (14.3)
Daily	1 (11.1)	0 (0.0)	1 (5.0)	4 (13.8)	1 (9.1)	1 (9.1)	8 (9.5)
Weekly	1 (11.1)	1 (25.0)	0 (0.0)	0 (0.0)	0 (0.0)	1 (9.1)	3 (3.6)
Monthly (every 1, 6, … months)	2 (22.2)	1 (25.0)	0 (0.0)	0 (0.0)	0 (0.0)	2 (18.2)	5 (6.0)
Yearly (every 1, 2, … years)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	1 (9.1)	1 (1.2)
At routine follow-up	5 (55.6)	2 (50.0)	15 (75.0)	16 (55.2)	8 (72.7)	6 (54.5)	52 (61.9)
Not fixed	0 (0.0)	0 (0.0)	1 (5.0)	0 (0.0)	0 (0.0)	0 (0.0)	1 (1.2)
Not reported	0 (0.0)	0 (0.0)	1 (5.0)	1 (3.45)	0 (0.0)	0 (0.0)	2 (2.4)
Prediction task ($N=124$)	($N=9$)	($N=4$)	($N=32$)	($N=48$)	($N=13$)	($N=18$)	($N=124$)
Clinical predictions	5 (55.5)	4 (100)	13 (40.6)	9 (18.8)	8 (61.5)	15 (83.3)	54 (43.5)
Cancer (eg, colorectal, pancreatic)	1 (11.1)	0 (0.0)	1 (3.12)	0 (0.0)	2 (15.4)	1 (5.6)	5 (4.0)
Cardiovascular system (eg, heart failure)	3 (33.3)	2 (50.0)	4 (12.5)	2 (4.17)	2 (15.4)	3 (16.7)	16 (12.9)
Infections (eg, sheptic shock)	0 (0.0)	0 (0.0)	2 (6.3)	3 (6.3)	2 (15.4)	0 (0.0)	7 (5.7)
Mental health (eg, depression, suicidal ideation)	0 (0.0)	1 (25.0)	2 (6.3)	0 (0.0)	0 (0.0)	0 (0.0)	3 (2.4)
Metabolic (eg, diabetes, obesity)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	6 (33.3)	6 (4.8)
Neurorological system (eg, Alzheimer's)	0 (0.0)	1 (25.0)	0 (0.0)	0 (0.0)	0 (0.0)	1 (5.6)	2 (1.6)
Respiratory system (eg, COPD)	0 (0.0)	0 (0.0)	2 (6.3)	4 (8.3)	1 (7.7)	3 (16.7)	10 (8.1)
Urinary system (eg, kidney disease)	1 (11.1)	0 (0.0)	2 (6.3)	0 (0.0)	1 (7.7)	1 (5.6)	5 (4.0)
Disease progression and health status	1 (11.1)	0 (0.0)	4 (12.5)	12 (25.0)	2 (15.4)	0 (0.0)	19 (15.3)
Decompensation	0 (0.0)	0 (0.0)	0 (0.0)	3 (6.3)	0 (0.0)	0 (0.0)	3 (2.4)
Mortality	1 (11.1)	0 (0.0)	4 (12.5)	9 (18.8)	2 (15.4)	0 (0.0)	16 (12.9)
Outcome measures for quality care	0 (0.0)	0 (0.0)	4 (12.5)	8 (16.7)	0 (0.0)	0 (0.0)	12 (9.7)
Hospital (re)admission	0 (0.0)	0 (0.0)	3 (9.4)	1 (2.1)	0 (0.0)	0 (0.0)	4 (3.2)
In-hospital mortality	0 (0.0)	0 (0.0)	0 (0.0)	4 (8.3)	0 (0.0)	0 (0.0)	4 (3.2)
Length of stay	0 (0.0)	0 (0.0)	1 (3.1)	3 (6.3)	0 (0.0)	0 (0.0)	4 (3.2)
Other predictions	3 (33.3)	0 (0.0)	11 (34.3)	19 (39.5)	3 (23.1)	3 (16.7)	39 (31.5)
Next event (eg, diagnose, drug)	3 (33.3)	0 (0.0)	10 (31.2)	15 (31.2)	2 (15.4)	1 (5.6)	31 (25.0)
Others (Freq. <2)	0 (0.0)	0 (0.0)	1 (3.1)	4 (8.3)	1 (7.7)	2 (11.1)	8 (6.5)
Prediction window ($N=117$)							
Hours (1, 3, 6, 8 h)	0 (0.0)	0 (0.0)	3 (8.3)	4 (12.1)	3 (18.8)	0 (0.0)	10 (8.6)
Days (1, 2, 7, 15 days)	0 (0.0)	0 (0.0)	8 (22.2)	7 (21.2)	0 (0.0)	1 (7.1)	16 (13.7)
Months (1, 2, 3, 6, 9 months)	2 (14.3)	0 (0.0)	12 (33.3)	4 (12.1)	3 (18.8)	4 (28.6)	25 (21.4)
Years (1, 2, 3, 4, 5, 10 years)	8 (57.1)	2 (50.0)	7 (19.4)	0 (0.0)	4 (25.0)	4 (28.6)	25 (21.4)
Any	4 (28.6)	2 (50.0)	6 (16.7)	18 (54.5)	6 (37.5)	5 (35.7)	41 (35.0)
Architecture							
RNN-based only	5 (71.4)	3 (75.0)	14 (70.0)	14 (50.0)	8 (72.7)	6 (54.5)	50 (61.7)
RNN/BiRNN	0 (0.0)	1 (25.0)	0 (0.0)	0 (0.0)	1 (9.1)	0 (0.0)	2 (2.5)
GRU/BiGRU	4 (57.1)	0 (0.0)	8 (40.0)	4 (14.3)	3 (27.3)	3 (27.3)	22 (27.2)
LSTM/BiLSTM	1 (14.3)	2 (50.0)	6 (30.0)	10 (35.7)	4 (36.4)	3 (27.3)	26 (32.1)
Transformer-based							
BERT-based architectures	0 (0.0)	0 (0.0)	1 (5.0)	2 (7.1)	2 (18.2)	2 (18.2)	7 (8.6)
Combinations	1 (14.3)	1 (25.0)	5 (25)	11 (39.3)	1 (9.1)	3 (27.3)	22 (27.2)
Variational RNN	0 (0.0)	0 (0.0)	1 (5.0)	0 (0.0)	0 (0.0)	0 (0.0)	1 (1.2)
CNN only or LSTM/GRU+CNN	0 (0.0)	1 (25.0)	4 (20.0)	4 (14.3)	0 (0.0)	3 (27.3)	12 (14.8)
DAG+GRU	1 (14.3)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	1 (1.2)
Dense only or LSTM/GRU+Dense	0 (0.0)	0 (0.0)	0 (0.0)	1 (3.6)	1 (9.1)	0 (0.0)	2 (2.5)
GAN+LSTM	0 (0.0)	0 (0.0)	0 (0.0)	1 (3.6)	0 (0.0)	0 (0.0)	1 (1.2)
GNN only or LSTM/GRU+GNN	0 (0.0)	0 (0.0)	0 (0.0)	4 (14.3)	0 (0.0)	0 (0.0)	4 (4.94)
GRU+GCN	0 (0.0)	0 (0.0)	0 (0.0)	1 (3.6)	0 (0.0)	0 (0.0)	1 (1.2)
Machine learning	1 (14.3)	0 (0.0)	0 (0.0)	1 (3.6)	0 (0.0)	0 (0.0)	2 (2.4)
Lasso-SVM	1 (14.3)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	1 (1.2)
Gradient boosting tree mimic	0 (0.0)	0 (0.0)	0 (0.0)	1 (3.6)	0 (0.0)	0 (0.0)	1 (1.2)
Number of layers							
<3	6 (85.7)	2 (50.0)	6 (30.0)	6 (21.4)	4 (36.4)	3 (27.3)	27 (33.3)
3-5	0 (0.0)	0 (0.0)	2 (10.0)	9 (32.1)	3 (27.3)	1 (9.1)	15 (18.5)
6-10	0 (0.0)	0 (0.0)	2 (10.0)	2 (7.14)	1 (9.1)	2 (18.2)	7 (8.6)
Not reported	1 (14.3)	2 (50.0)	10 (50.0)	11 (39.3)	3 (27.3)	5 (45.5)	32 (39.5)

Categorical parameters are described as N (%). Some categories have been aggregated. Raw parameters are available in the Shiny app. $N>81$ is due to the same study considering different possibilities for the same parameter.

Table 3. Description of the parameters related to the development of the model.

Parameter	Description
Handling of missing values	26/81 (32.1)
Compatible with model	4/26 (15.4)
Imputation	17/26 (65.4)
Not reported	5/26 (19.2)
Preprocessing	78/81 (96.3)
One-hot encoding	45/78 (57.7)
Embedding	42/78 (53.8)
Time window aggregation	26/78 (33.3)
Simultaneous preprocessing techniques (*n*)	[0, 5], 1.9 (1), 2 [1-3], 100
Feature selection	6 (7.4)
Varying-length sequence handling	
Preprocessing	31/81 (38.3)
Zero-padding	22/81 (27.2)
Not reported	19/81 (23.5)
Not needed	5/81 (6.2)
Number of layers	[0,10], 2.9 (2), 2 [2-4], 60.5
Use of attention mechanism	45/81 (55.6)
Use of static variables	27/81 (33.3)
Hyperparameter tuning	
Performed	36/81 (44.4)
Not performed	34/81 (42)
Not reported	11/81 (13.6)
Hyperparameter tuning method	
Fine-tuning	13/36 (36.1)
Grid search	7/36 (19.4)
Others	10/36 (27.8)
Not reported	6/36 (16.7)
Hyperparameters tuned	
Number of neurons per layer	16/36 (44.4)
Learning rate	10/36 (27.8)
Dropout rate	8/36 (22.2)
Simultaneous hyperparameters being tuned (*n*)	[1, 8], 2.6 (1.9), 2 [1-3.3] (80.6)

Categorical parameters are described as *N* (%), while quantitative parameters as [min, max], mean (SD), median [Q1-Q3], (% studies reported).

Table 4. Description of the parameters related to the model training.

Parameter	Description
Regularization mechanisms	52/81 (64.2)
Dropout	41/52 (78.9)
L1 or L2 regularization	23/52 (45.1)
Others	21/52 (40.4)
Simultaneous regularization techniques (*n*)	[1, 5], 1.6 (0.8), 1 [1-2]
Optimizer	
Adam	39/81 (48.1)
Stochastic gradient descent (SGD)	5/81 (6.2)
Adadelta	5/81 (6.2)
Others	7/81 (8.6)
Not reported	25/81 (30.9)
Internal validation	
Random split	54/81 (66.7)
Cross-validation	15/81 (18.5)
Others	3/81 (3.7)
Not reported	9/81 (11.1)
Training set size (%)	[14, 90], 73.8 (10.7), 75 [70-80], (92.6)
Validation set size (%)[a]	[4.4, 50], 14.5 (7.5), 10 [10-16.9], (16.1), (12.3)
Test set size (%)[a]	[10, 80], 18.1 (9.5), 16.7 [15-20], (6.2), (16.1)
Measure of performance variability[b]	44/81 (54.3)
External validation	3/81 (3.7)
Comparison with	
Simpler models	70/81 (86.4)
State-of-the-art models	37/81 (45.7)

Categorical parameters are described as *N* (%), while quantitative parameters as [min, max], mean (SD), median [Q1-Q3], (% studies reported).

[a] Stands for the percentages of studies that did not report the information and that did not define those sets.

[b] Use of techniques like cross-validation or bootstrapping to quantify the variability of the performance of the model in the internal validation.

(Table 2). However, in studies with regular assessment points, the more severe the patient's condition, the more frequent the data registers were (eg, 27.6% ICU studies had hourly time points, while 22.2% of primary care studies had a monthly register). This difference in recording frequency plus the total length of available follow-up in each setting affected the number of registers per patient. In inpatient and ICU settings, the number was usually 10 or less, compared to 51-100 in primary care. Developing particular diseases, mainly cardiovascular diseases, and the next event were the most common prediction tasks in all settings. In addition, 25% of ICU studies aimed to predict disease progression and health status (eg, decompensation, mortality). Hospitalization and ICU prediction windows were shorter (ie, hours, days, or months) than most outpatient care (ie, primary care, consultations), which used years as the prediction window.

Regarding the architecture, most studies considered only RNN, GRU, or LSTM layers regardless of data context. However, several ICU studies also combined these layers with others, such as CNN or GNN. As a result, ICU studies tended to use more layers (ie, most reported using 3 or more layers, while the architectures used in the rest of the settings were based on fewer layers). Therefore, the number of registers available per patient and the frequency of registers can modify the choice of architecture.

Performance

Reporting of performance metrics was highly variable. We collected the following where reported: precision, accuracy, recall/sensitivity, specificity, F_2-score, AUC, and AUC precision-recall. AUC refers to the area under the sensitivity versus (1-specificity) curve, or ROC curve, while AUC precision-recall refers to the area under the precision versus sensitivity curve. The included studies reported 0-5 of these metrics (mean 1.83, SD 1.10), with most reporting just 1 (33/81, 40.7%) or 2 (28/81, 34.6%). The metrics most commonly reported were AUC (58/81, 71.6%), followed by AUC precision-recall (average precision; 29/81, 35.8%) and precision (22/81, 27.2%). AUC varied from 0.75 to 0.99 (mean 0.86, SD 0.07), and AUC precision-recall from 0.13 to 0.87 (mean 0.54, SD 0.20). A subanalysis of the performance obtained by the best model reported in the studies using MIMIC-III is reported in Table 5.

Completeness of reporting assessment—TRIPOD statement

Figure 2 shows the adherence to the TRIPOD subitems. Overall adherence to the TRIPOD statement ranged from 35.1%

Table 5. Performance metrics of the best models reported by the studies using the MIMIC-III dataset (*N* = 21), by outcome and method (simplified).

Outcome	Method	AUC	AUC precision-recall
Disease progression and health status (*N* = 9)	BERT-based architectures (*N* = 1)	0.859 (—)	0.519 (—)
	RNN-based only (*N* = 5)	0.868 (0.063)	0.496 (0.068)
	Combinations (*N* = 3)	0.901 (0.001)	0.326 (0.023)
Outcome measures for quality care (*N* = 3)	RNN-based only (*N* = 3)	0.846 (0.025)	0.525 (0.007)
Next event (diagnose, drug, etc.) (*N* = 9)	BERT-based architectures (*N* = 1)	—	0.700 (—)
	RNN-based only (*N* = 3)	0.921 (0.038)	0.835 (0.049)
	Combinations (*N* = 5)	0.839 (—)	0.504 (0.228)

Metrics described as mean (SD). — denotes that the aggregation could not be computed, either because it was not reported or it was reported only by one study. Only AUC and AUC precision-recall are reported, as they were the metrics most reported in the studies.

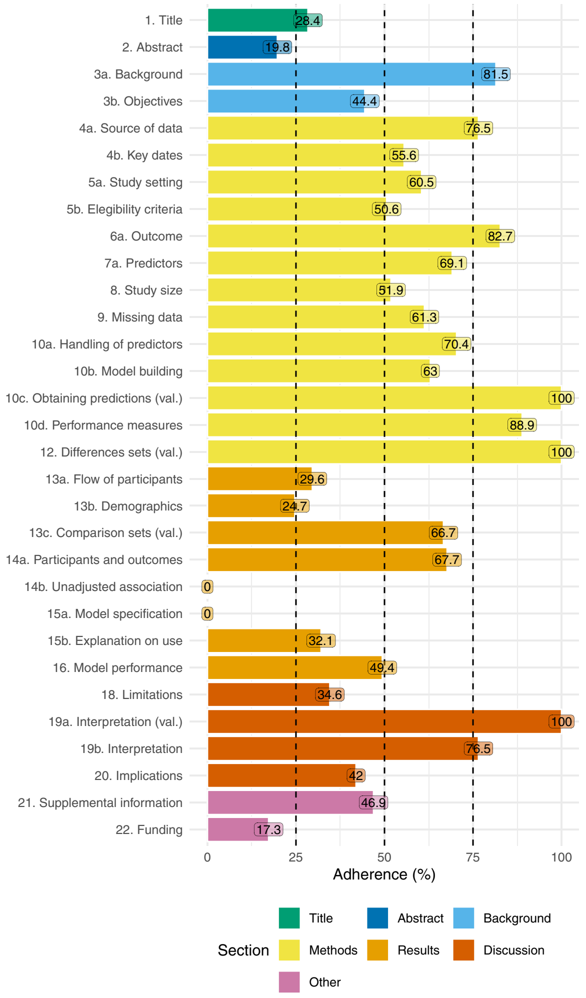

Figure 2. Assessment of the items of the TRIPOD statement. Overall sample *n* = 81. Out of the 37 items, 6 (ie, 5c, 6b, 7b, 10e, 11, 17) are not reported as they were not applicable to any of the studies considered.

to 89.2% (median 62.16%, IQR [54.1%, 73%]). The title, abstract, and introduction were appropriately reported in 23/81 (28.40%), 16/81 (19.8%), and 31/81 (38.3%) studies, respectively. In contrast, only 5/81 (6.2%) appropriately reported all subitems in the "Methods" section, 0/81 (0%) in the "Results" section, and 16/81 (19.8%) in the "Discussion" section. Other information was adequately reported in 8/81 (9.9%) studies. Those using transformer-based models, whose methodology may be the most complex, showed the highest adherence to TRIPOD (see Supplementary Appendix F).

Risk of bias assessment—PROBAST statement

The short form of PROBAST assessed the outcome and the analysis domains and was applied in 22/81 studies that reported the necessary information (see Supplementary Appendixes C and D). Three items regarding the outcome domain were assessed: the determination of the outcome was or probably was appropriate in 19/22 (86.4%) models (item 3.1), prespecified or standard definitions were used in 17/22 (77.3%) models (item 3.2), and it was defined and determined similarly for all participants in 21/22 (95.5%) models (Figure 3). Thus, 72.7% of the reviewed models had a low risk of bias in the outcome domain. Regarding the analysis domain, 5 items were assessed: there were a reasonable number of participants with the outcome in 16/22 (72.7%) models (item 4.1), the variables were handled appropriately in 17/22 (77.3%) (item 4.2), the information to assess missing handling was insufficient in 10/22 (45.5%) models (item 4.4), univariable selection was avoided in 20/22 (90.9%) models (item 4.5), and the overfitting and optimism were considered in 15/22 (68.2%) models. Altogether, the information to assess the risk of bias in the analysis domain was insufficient in 11/22 (50%) models. Considering both domains, the overall risk of bias was high in 7/22 (31.8%) models, unclear in 12/22 (54.6%), and low in 3/22 (13.6%). The studies using architectures based on RNN-like layers only were at the highest risk of bias (see Supplementary Appendix F).

Discussion

Key results

We included 81 studies and extracted the following key results. Regarding the data specifications, more than a third of the studies used ICU data, generally offering short follow-up times but with high temporal granularity. We expected more studies set in primary care, as it is the context that offers more longitudinality. Data sampling frequency was very heterogeneous depending on the context, varying from hourly to every several years. The different amounts of data per patient and the time spacing between them influenced the choice of architecture, increasing its complexity as less spaced in time were the registers. Almost half the prediction tasks were focused on the development of certain conditions or the next events. Very few studies included socioeconomic characteristics in the models, despite their importance as health determinants.[32–34]

Regarding missing values, researchers need to make more efforts to explain how they dealt with them, as this aspect required clarification in most studies. Models that can handle

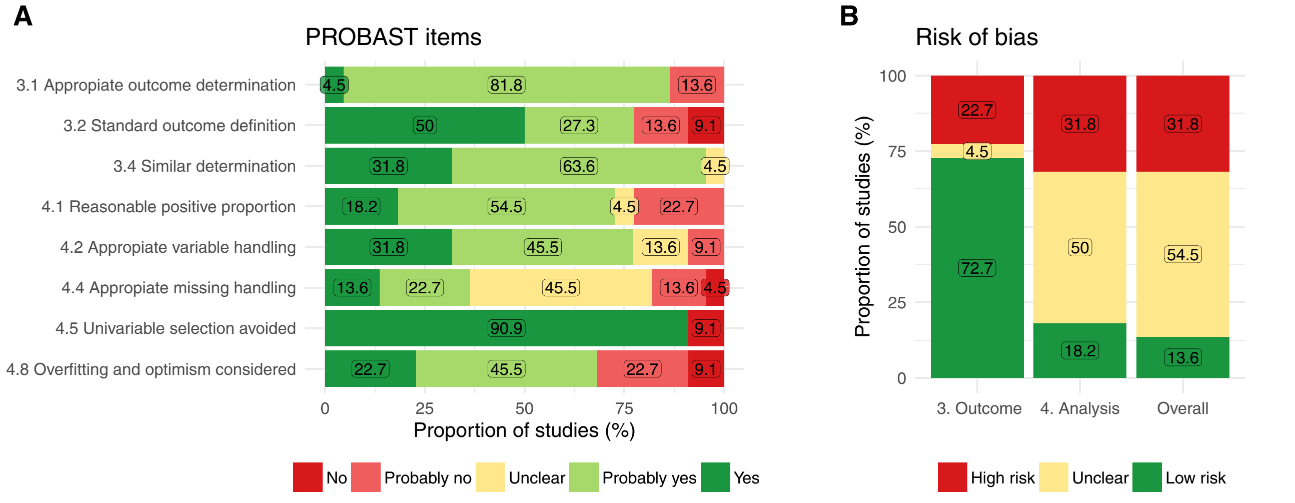

Figure 3. Assessment of PROBAST items, as proposed by Venema et al.[29] Sample $n = 22$. (A) Reports each item individually and (B) reports risk of bias overall and by section. The numbering of the items and domains refers to the original PROBAST.[31]

missing data without imputation (eg, masking) should be considered more often. Conversely, most studies did some preprocessing of the data and reported it correctly. Few studies changed the actual target distribution in training or made feature selection, which can be considered positive. The number of layers used is lower than in other applications such as computer vision.[35] However, the number of registers available per patient and their frequency modified the architecture; the more registers per person, the more layers and the more types of layers used. For example, the studies with ICU data (ie, less absolute follow-up time per patient, but more frequent registers) used more complex architectures in terms of the number and types of layers involved. Around half the studies incorporated attention mechanisms to increase either model performance or explainability. The use of these mechanisms in healthcare is of great interest, as patients and professionals may need to understand the reasoning for the prediction. Hyperparameter tuning improves model performance, but only about half the studies used it, probably due to its computational cost. There also is room for improvement in model validation, as most studies performed only an internal validation with a simple train/test split, without estimating the variance of the model's performance, and external validation was performed in just 3 studies. This could also be related to the computational cost of cross-validating on large datasets or the lack of available public datasets that could be used as benchmarks.

According to the subanalysis on the studies using MIMIC-III, models considering longitudinal data achieved reasonably good results. When predicting disease progression and health status, eg, death or decompensation, all methods produced good results. Using a combination of different types of layers (recurrent and nonrecurrent) achieved the highest AUC, but the better balance between classes, ie, the higher AUC precision-recall, was achieved using BERT-based architectures. Regarding the prediction of next events, eg, next prescription, architectures using only RNN-like layers, eg, LSTM or GRU, achieved both the higher AUC and AUC precision-recall.

The overall quality of the reporting was not optimal (median adherence: 62%). It was even more difficult to find a study that reported all items appropriately. Andaur-Navarro[36] also described this situation, which is common for other study designs like randomized controlled trials.[37] Risk of bias came mainly from the analysis section, which was not properly reported in 50% of the studies assessed, and 31.6% had a high overall risk of bias. Similar results were reported by Andaur-Navarro.[38] The studies that built their models using more complex architectures (eg, transformers, combinations of different types of layers) had better reporting and less risk of bias than studies following simpler approaches.

Comparison with other studies

This systematic methodological review identified and analyzed different AI-based techniques used to predict health-related outcomes based on longitudinal data from EHRs. Previous works in this field that broadly identified studies using ML in health also reported a lack of homogeneity in reporting, and a limited use of primary care data.[21] Silva et al[8] reported different technical approaches to longitudinal data in DL architectures, but not through a systematic review.

Previous systematic reviews of supervised ML in health have also described poor reporting and high risk of bias,[36,38] calling for a better reporting quality and better explanations of the approach to missing values.[39] Poor reporting and the lack of data and code sharing hampers the reproducibility of prediction models.[40]

Strength and limitations of this study

The main strength of this review is that it provides an extensive overview of state of the art in AI-based models using longitudinal EHR data and some recommendations for modeling with this type of data. In addition, it provides an online interactive dashboard to further explore the collected parameters. However, this systematic review has several limitations. Firstly, we excluded short reports, and although this type of publication generally does not include sufficient information[24–26] to be able to accurately assess all aspects assessed in this review, there is a possibility that some relevant studies could be excluded. Secondly, we used comprehensive study search strategies that combined a wide variety of search terms from free text and subject headings (see Supplementary Appendix A). However, these search strategies did not include some search terms based on single concepts to maximize sensitivity whilst striving for some reasonable precision.[41] For example, using the single term “longitudinal” instead of the pairs of terms we used (eg, “longitudinal data,” “longitudinal study,” etc.), could have retrieved a larger number of studies that were longitudinal by design but that did not necessarily analyze the data longitudinally. Nevertheless, this approach to search strategies may have contributed to some relevant reports not being identified. Therefore, to reduce the potential risk of publication bias introduced by these 2 limitations and to identify as many relevant studies as possible, we supplemented this identification by checking the reference lists of the included studies and consulting with experts in the field. This search of other sources allowed us to identify 32 additional reports (see Figure 1). Finally, using the short-form PROBAST instead of the original may be a limitation. However, this was necessary to obtain more detail on the risk of bias, as the type of data used (ie, EHR) would already confer a high risk of bias on all the included studies. In this line, it is necessary to develop risk of bias tools that can appropriately assess studies based on EHRs, which, if used correctly, should not be considered an inherent source of bias.[42] Even though a quantitative analysis was planned in the review protocol, it was not possible to carry it out considering all the included studies due to the heterogeneity in the reporting of performance metrics, and it was performed only on a subset of them using MIMIC-III. This comparison is limited as each study could have used a different subset of subjects, but it was performed assuming they have the same measurement bias, type of missingness, and underlying patient population.

Implication for researchers, editorial offices, and future studies

Most studies focused on predicting the development of conditions or the occurrence of certain events (eg, death, drug prescription), while only 9.7% predicted outcomes related to quality of care or management. More efforts in this direction could help inform resource planning.

Studying model performance according to the technique used was difficult due to the heterogeneity in the metrics reported, precluding any recommendations on which architecture best suited the different data types. In the absence of any consensus, one recommendation that does emerge is for future studies to report as many performance metrics as possible, to facilitate the development of benchmarks and enable the analysis of the correct identification of both positive and negative cases, which cannot be distilled if only one metric, like accuracy, is reported. The original distribution of the target (ie, class imbalance) should be considered when interpreting the metrics in order not to assume that a model is good simply because it has a good AUC. An extensively defined set of prediction tasks for models in health that could be used for benchmarking could also be proposed, since we also found a high variability in the definition of outcomes that made it difficult to compare studies. Defining such standards or benchmarks is particularly important in an area that is evolving as rapidly such as AI to facilitate understanding, comparability, and reproducibility.

Furthermore, a notably high number of studies used ICU data to build their longitudinal models. This may be due to MIMIC-III, a database of ICU data, being one of the few publicly available databases including real EHR data. The availability of other large databases that include EHRs from other data contexts (eg, primary care) is vital to develop more diverse prediction models. If more studies had used primary care data, there would probably be more studies focusing on quality-of-care indicators, such as hospital admissions. Therefore, we encourage researchers to make their databases available in some form. Sharing both data and the code that built and trained the models, following open science principles,[43] would improve the replicability of the studies. If the data cannot be shared, codes could be published with a minimal set of simulated data to run the code, creating a minimal, reproducible example.

Regarding the risk of bias, the treatment of missingness should be improved, more efforts made to study the presence of overfitting and consider optimism in model performance. Developing guidelines for properly assessing both reporting and risk of bias in studies using AI is necessary and on its way.[30] In light of this study, we propose that the forthcoming guidelines encompass the following considerations:

- Thoroughly report the handling of data heterogeneity, specifically in terms of missingness and length of follow-up. This is particularly crucial if the model itself addresses these issues rather than relying on preprocessing steps.
- Provide a comprehensive set of metrics that adequately evaluate the model’s performance, taking into account the aforementioned data heterogeneity.
- Offer detailed documentation on the architecture and implementation of the model. Transparency regarding the number of layers, utilization of masking or normalization layers, and the tuning of hyperparameters should not be obscured or undisclosed. Sharing analytical code would help with this. In addition, it would foster a culture of continuous improvement that accelerates progress and ensures research’s quality and reproducibility.

Conclusion

This review found that AI-based models using longitudinal EHR data to predict health- or healthcare-related outcomes are being developed in different healthcare settings, using mainly information related to diagnoses and drugs. RNN-based architectures are the most common approach when considering longitudinal data, but transformers and other combinations of layers are also being used. Most models are trained and validated using a simple train-test split, but only about half measured the performance variability to estimate optimism and overfitting. There is a significant lack of homogeneity when reporting both methodology and performance, complicating the comparison of the results achieved by these models. The overall quality of the reporting was rated at just 62%, and 31.6% of the assessed studies had a high risk of bias, underscoring the need for the development of reporting guidelines for this kind of studies. AI models that are capable of considering temporal information are an innovative improvement in biomedical and health informatics research, but researchers should also mind how they report their work, and not just on improving techniques, to ensure that the results are beneficial to the wider scientific community.

Acknowledgments

The authors thank Meggan Harris for her role in editing and proofreading the manuscript.

Ethical approval

Not required.

Author contributions

L.A.C.-R., E.Z.-O., and C.V. participated in the conceptualization of the study. L.A.C.-R., M.M.-Z., E.Z.-O., M.C.-B., and C.V. screened the articles. L.A.C.-R., J.L.-J., and C.G.-M. performed the full article examination and extracted the data. L.A.C.-R. performed the formal data analysis and wrote the first draft of the manuscript. All the authors participated in the review and editing of the final manuscript.

Supplementary material

Supplementary material is available at *Journal of the American Medical Informatics Association* online.

Funding

The project received a research grant from the Carlos III Institute of Health, Ministry of Economy and Competitiveness (Spain), awarded on the 2019 call under the Health Strategy Action 2013-2016, within the National Research Programme oriented to Societal Challenges, within the Technical, Scientific and Innovation Research National Plan 2013-2016 (reference PI19/00535), and the PFIS Grant FI20/00040, co-funded with European Union ERDF (European Regional Development Fund) funds. The project has also been partially funded by Generalitat de Catalunya through the AGAUR (grant numbers 2021-SGR-01033, 2021-SGR-01537). The funders had no role in the study design, data collection, data analysis, data interpretation, or writing of this work.

Conflicts of interest

None declared.

Data availability

The authors confirm that the data supporting the findings of this methodological systematic review are available within the article and its Supplementary Material. The parameters collected can be directly downloaded from our Shiny app (https://lacarrascoribelles-idiapjgol.shinyapps.io/SR_AILongitudinalModels/).

References

1. Liu P-R, Lu L, Zhang J-Y, Huo T-T, Liu S-X, Ye Z-W. Application of artificial intelligence in medicine: an overview. *Curr Med Sci*. 2021;41(6):1105-1115.
2. Mintz Y, Brodie R. Introduction to artificial intelligence in medicine. *Minim Invasive Ther Allied Technol*. 2019;28(2):73-81.
3. Shillan D, Sterne JAC, Champneys A, Gibbison B. Use of machine learning to analyse routinely collected intensive care unit data: a systematic review. *Crit Care*. 2019;23(1):284.
4. Buchlak QD, Esmaili N, Leveque J-C, et al. Machine learning applications to clinical decision support in neurosurgery: an artificial intelligence augmented systematic review. *Neurosurg Rev*. 2019;43(5):1235-1253.
5. James MT. Longitudinal studies 4: matching strategies to evaluate risk. In: Parfrey PS, Barrrett BK, eds. *Clinical Epidemiology. Methods in Molecular Biology*. Vol. 2249. Springer; 2021:167-177.
6. Gaspar PM, Bautch JC, Strodthoff SCM. A longitudinal study of the health status of a community of religious sisters: addressing the advantages, challenges, and limitations. *Res Gerontol Nurs*. 2015;8(2):77-84.
7. Konerman MA, Zhang Y, Zhu J, Higgins PDR, Lok ASF, Waljee AK. Improvement of predictive models of risk of disease progression in chronic hepatitis C by incorporating longitudinal data. *Hepatology* 2015;61(6):1832-1841.
8. Silva JF, Matos S. Patient trajectory modelling in longitudinal data: a review on existing solutions. In: *2021 IEEE 34th International Symposium on Computer-Based Medical Systems (CBMS)*. IEEE; 2021; Aveiro, Portugal.
9. Dorr DA, Ross RL, Cohen D, et al. Primary care practices' ability to predict future risk of expenditures and hospitalization using risk stratification and segmentation. *BMC Med Inform Decis Mak*. 2021;21(1):104.
10. Beau Hilton C, Milinovich A, Felix C, et al. Personalized predictions of patient outcomes during and after hospitalization using artificial intelligence. *NPJ Digit Med*. 2020;3(1):51.
11. Bernardini M, Romeo L, Frontoni E, Amini M-R. A Semi-Supervised Multi-Task learning approach for predicting short-term kidney disease evolution. *IEEE J Biomed Health Inform*. 2021;25(10):3983-3994.
12. Allam A, Feuerriegel S, Rebhan M, Krauthammer M. Analyzing patient trajectories with artificial intelligence. *J Med Internet Res*. 2021;23(12):e29812.
13. Hochreiter S, Schmidhuber J. Long short-term memory. *Neural Comput*. 1997;9(8):1735-1780.
14. Cho K, van Merrienboer B, Gulcehre C, et al. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics; 2014. https://doi.org/10.3115/v1/d14-1179
15. Ma L, Gao J, Wang Y, et al. AdaCare: explainable clinical health status representation learning via scale-adaptive feature extraction and recalibration. *Proc AAAI Conf Artif Intell* 2020;34(01):825-832.
16. An Y, Tang K, Wang J. Time-aware multi-type data fusion representation learning framework for risk prediction of cardiovascular diseases. *IEEE/ACM Trans Comput Biol Bioinform*. 2021;19(6):3725-3734.
17. Luong M-T, Pham H, Manning CD. Effective approaches to attention-based neural machine translation. *arXiv:1508.04025v5*. 2015.
18. Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate. *arXiv:1409.0473v7*. 2016.
19. Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. *arXiv:1706.03762v7*. 2017.
20. Li Y, Rao S, Solares JRA, et al. BEHRT: transformer for electronic health records. *Sci Rep*. 2020;10(1):7155.
21. Andaur Navarro CL, Damen JA, van Smeden M, 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.
22. Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ* 2021;372:n71.
23. Munn Z, Stern C, Aromataris E, Lockwood C, Jordan Z. What kind of systematic review should I conduct? A proposed typology and guidance for systematic reviewers in the medical and health sciences. *BMC Med Res Methodol*. 2018;18(1):5.
24. Mayo-Wilson E, Li T, Fusco N, Dickersin K; MUDS Investigators. Practical guidance for using multiple data sources in systematic

reviews and meta-analyses (with examples from the MUDS study). *Res Synth Methods*. 2017;9(1):2-12.

25. Li T, Higgins JPT, Deeks JJ. Chapter 5: collecting data. In: Higgings JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, Welch VA, eds. *Cochrane Handbook for Systematic Reviews of Interventions*. Version 6.3. Cochrane; 2022.
26. Hopewell S, Clarke M, Askie L. Reporting of trials presented in conference abstracts needs to be improved. *J Clin Epidemiol*. 2006;59(7):681-684.
27. Ouzzani M, Hammady H, Fedorowicz Z, Elmagarmid A. Rayyan—a web and mobile app for systematic reviews. *Syst Rev*. 2016;5(1):210.
28. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *Ann Intern Med*. 2015;162(1):55-63.
29. Venema E, Wessler BS, Paulus JK, et al. Large-scale validation of the prediction model risk of bias assessment tool (PROBAST) using a short form: high risk of bias models show poorer discrimination. *J Clin Epidemiol*. 2021;138:32-39.
30. Collins GS, Dhiman P, Andaur Navarro CL, et al. 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):e048008.
31. Wolff RF, Moons KGM, Riley RD, et al.; PROBAST Group. PROBAST: a tool to assess the risk of bias and applicability of prediction model studies. *Ann Intern Med*. 2019; 170(1):51-58.
32. Nihtilä E, Martikainen P. Why older people living with a spouse are less likely to be institutionalized: the role of socioeconomic factors and health characteristics. *Scand J Public Health*. 2008;36(1):35-43.
33. Stamatakis E, Primatesta P, Chinn S, Rona R, Falascheti E. Overweight and obesity trends from 1974 to 2003 in english children: what is the role of socioeconomic factors? *Arch Dis Child*. 2005;90(10):999-1004.
34. Braveman PA, Heck K, Egerter S, et al. The role of socioeconomic factors in Black–White disparities in preterm birth. *Am J Public Health*. 2015;105(4):694-702.
35. Yoo H-J. Deep convolution neural networks in computer vision: a review. *IEIE Trans Smart Process Comput*. 2015;4(1):35-43.
36. Navarro CLA, Damen JAA, Takada T, et al. Completeness of reporting of clinical prediction models developed using supervised machine learning: a systematic review. *BMC Med Res Methodol*. 2022;22(1):12.
37. Shahzad R, Ayub B, Siddiqui MAR. Quality of reporting of randomised controlled trials of artificial intelligence in healthcare: a systematic review. *BMJ Open* 2022;12(9):e061519.
38. Andaur Navarro CL, Damen JAA, Takada T, et al. Risk of bias in studies on prediction models developed using supervised machine learning techniques: systematic review. *BMJ* 2021; 375:n2281.
39. Nijman SWJ, Leeuwenberg AM, Beekers I, et al. Missing data is poorly handled and reported in prediction model studies using machine learning: a literature review. *J Clin Epidemiol*. 2022;142:218-229.
40. Belbasis L, Panagiotou OA. Reproducibility of prediction models in health services research. *BMC Res Notes*. 2022;15(1):204.
41. Lefebvre C, Glanville J, Briscoe S, et al. Chapter 4: searching for and selecting studies. In: Higgings JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, Welch VA, eds. *Cochrane Handbook for Systematic Reviews of Interventions*. Version 6.3. Cochrane; 2022.
42. Sauer CM, Chen L-C, Hyland SL, Girbes A, Elbers P, Celi LA. Leveraging electronic health records for data science: common pitfalls and how to avoid them. *Lancet Digit Health*. 2022;4(12):e893-e898.
43. UNESCO. Records of the general conference: 41st session. In: United Nations Educational, Scientific and Cultural Organization, ed. *Resolutions, Chapter Annex VI: Recommendation on Open Science*. United Nations Educational; 2022:137-150.