Predicting depression using electronic health records data: A systematic review

David Nickson (david.nickson@warwick.ac.uk)
University of Warwick
Caroline Meyer
University of Warwick
Lukasz Walasek
University of Warwick
Carla Toro
University of Warwick

Research Article

Keywords: Artificial Intelligence, Depression, Diagnosis, Electronic Health Records, Machine Learning, Prediction

Posted Date: February 1st, 2023

DOI: https://doi.org/10.21203/rs.3.rs-2510168/v1

Abstract

Background

Depression is one of the most significant health conditions in personal, social, and economic impact. The aim of this review is to summarize existing literature in which machine learning (ML) methods have been used in combination with Electronic Health Records (EHRs) for prediction of depression.

Methods

Systematic literature searches were conducted within arXiv, PubMed, PsycINFO, Science Direct, SCOPUS and Web of Science electronic databases. Searches were restricted to information published after 2010 (from 1st January 2011 onwards) and were updated prior to the final synthesis of data (27th January 2022).

Results

Following the PRISMA process, the initial 744 studies were reduced to 19 eligible for detailed evaluation. Data extraction identified machine learning methods used, types of predictors used, the definition of depression, classification performance achieved, sample size, and benchmarks used. Area Under the Curve (AUC) values more than 0.9 were claimed, though the average was around 0.8. Regression methods proved as effective as more developed machine learning techniques.

Limitations

The categorization, definition, and identification of the numbers of predictors used within models was sometimes difficult to establish, Studies were largely Western Educated Industrialised, Rich, Democratic (WEIRD) in demography.

Conclusion

This review supports the potential use of machine learning techniques with EHRs for the prediction of depression. All the selected studies used clinically based, though sometimes broad, definitions of depression as their classification criteria. The reported performance of the studies was comparable to or even better than that found in primary care. There are concerns over the generalizability and interpretability.

Background

Depression is the most common mental health condition globally, with one-year global prevalence rates ranging from 7 to 21% [1]. Quality of life can be seriously impaired by this disorder, with depression ranking as the second highest cause of Disability-Adjusted Life Years (DALYs) and Years Lived with Disability (YLDs) [2,3]. Depression is a major contributory factor in suicide affecting hundreds of thousands of cases per year [4,5]. In addition to the significant personal and social impact of depression, there is a significant economic cost. For example, in 2007 alone, total annual costs of depression in England were £7.5 billion, of which health service costs comprised £1.7 billion and lost earnings £5.8 billion [6,7].

Depression, like most mental health disorders, can be difficult to diagnose, especially for non-specialist clinicians [8,9]. Assessment by primary or secondary care clinicians typically relies on the World Health Organisation's International Catalogue of Diseases version 10 or 11, ICD-10/11 [10], the Diagnostic and Statistical Manual of Mental Disorders DSM [11], or by using an interview script such as the Composite International Diagnostic Interview (CIDI) [12,13]. Diagnosis can also be aided by garnering self-reported symptoms in response to standardised questionnaires such as the Hospital Anxiety and Depression Scale (HADS) [14], Beck Depression Inventory (BDI) [15,16] and Patient Health Questionnaire-9 (PHQ-9) [17,18]. The PHQ-9 is considered a gold standard [19] for screening rather than standalone clinical diagnosis [20] and has been validated internationally [18]. As such it sets a sound benchmark for sensitivity (e.g., 0.92) and specificity (e.g., 0.78) that is a good comparator for assessing alternative methods [21].

Considering mental health care pathways, benefits to patients could be provided by early diagnosis, opening the possibility to early interventions. For example, Bohlmeijer *et al.* [22] observed reduced symptoms of depression for patients who engaged in acceptance and commitment therapy (ACT) as an early intervention compared to those on a wait list, both initially and at a three month follow up. Furthermore, a meta-analysis by Davey and McGorry [23] showed a reduction in the incidence of depression by about 20% in the 3 to 24 months following an early intervention. At the same time, late diagnoses of depression can result in longer term suffering for the patient in terms of symptoms experienced and disorder trajectory together with increased resource consumption [8,24].

Recently, attempts to support early medical diagnoses have benefited from a) growing availability of electronic healthcare records (EHRs) that contain patients' longitudinal medical histories and b) new advances in predictive modelling and machine learning (ML) approaches. The use of EHRs in primary care in the developed world is well established. For example, in the USA, UK, Netherlands, Australia and New Zealand, take up in primary care has exceeded 90% [25,26]. The wide availability of proprietary EHR systems such as SNOMED in the UK [27] are enabling rapid and global implementation and their use for disorder surveillance [28]. For example, ML techniques with EHR data have led to predictive models for cardiovascular conditions [29,30] and diabetes [31]. These studies have led to cardiovascular risk prediction becoming established in routine clinical care and the UK QRISK versions 2 and 3 show significant improvements in discrimination performance over the Framingham Risk Score and atherosclerotic cardiovascular disease (ASCVD) score methods [32] that preceded them. A scoping review by Shatte *et al.* [33] on the general use of ML in mental health identified the use of ML with EHRs for identifying depression as a research area. Cho *et al.* [34] included depression amongst the conditions they identified in their "Review of Machine Learning Algorithms for Diagnosing Mental Illness". Investigating EHR/ML as a means of predicting depression diagnosis is a way forward.

If EHR/ML methods are to be considered, a suitable benchmark comparator is needed. Studies assessing diagnosis of depression in primary care suggest that approximately half of all cases are missed at first consultation but that this improves to around two thirds being diagnosed at follow up [35–37]. This would be a useful minimum comparator for any diagnostic system based on a combination of ML and EHRs data. There exists the potential to develop predictive models of depression using EHR/ML applications and it is necessary to critically evaluate how the field has evolved over the years. This is particularly important in the context of rapidly developing ML techniques, and the growing accessibility and richness of health data. Therefore, the objectives of this systematic review are to identify and evaluate studies that have used such techniques. As part of the evaluation, we specifically focus on identifying key features of the data and statistical methods used. Accordingly, our primary focus is to provide a comprehensive overview of the types of ML models and techniques used by researchers, as well as types of data on which these models were trained. By summarizing main properties of the data, identifying and summarising predictors used, describing diagnostic benchmarks, and outlining what types of validation approaches were used, our review offers an important source of information for those who wish to build on existing efforts to improve predictive accuracy of such models.

Methods

Search Strategy and Search Terms

Systematic literature searches were conducted within arXiv, PubMed, PsycINFO, Science Direct, SCOPUS and Web of Science electronic databases. Searches were restricted to information published after 2010 (from 1st January 2011 onwards) and were updated prior to the final synthesis of data on 27th January 2022. Initial searches were made based on titles/key words (where latter available) and papers were selected based on the inclusion criteria summarised in Table 1. These were searched as (#1) AND (#2) AND (#3) AND (#4). These papers were evaluated by reading the Abstract, and then by evaluating main body of each manuscript. Next, a backward citation search for all the selected papers was completed as both a) a quality check to see if other selected papers were included and b) to identify any missing papers. The last search step was a forward search pass where papers that cited the selected papers were identified; again, identifying any missed papers. The primary evaluation was conducted by DN and LW.

This systematic review was prospectively registered with Prospero international database of systematic reviews (# CRD42021269270) [38].

Table 1
Search terms for study identification

Component	*Area*	*Search terms*
#1	Artificial Intelligence/Machine Learning	(artificial intelligence) OR (machine learning) OR (data mining) OR (supervised learning) OR (unsupervised learning) OR (predictive analytics) OR (reinforcement learning) OR deep learning)
#2	Screening/Diagnosis	(screening, including: screen*; identif* detect*) OR (diagnosis including diagnos*) OR (Classification) OR (prediction including: predict*)
#3	Depression	Depression OR Depressive
#4	Electronic Health Records	(Electronic Health Records, including EHR) OR (Electronic Medical Records, including EMR) OR (Clinical records) OR Clinical notes)

Inclusion/Exclusion Criteria

Table 2 shows the inclusion and exclusion criteria that were adopted to define the publications that came within the scope of the review.

Table 2
Inclusion/Exclusion Criteria

Inclusion	*Exclusion*
Screening/Prediction/Diagnosis of depression in the undiagnosed with/without comorbidities	Involved interventions/trials or delivery/monitoring of interventions
Artificial Intelligence/Machine Learning techniques	Used additional unproven, experimental, bespoke or laboratory technology;
Used EHRs/Clinical notes derived data as primary source	Used additional high cost/specialist technology such as fMRI scanners, ECG, PET scans, radiography etc.
Ethically approved	Involved invasive procedures such as blood tests, CSF assays
Took place after 01/January/2011	Required additional activity to obtain predictor data e.g., clinical interviews.
Available in English	Review/Summary paper
Published in a peer reviewed journal/recognised publisher/conference paper.	

Data Extraction

Data extraction was informed by requirements detailed in: ‘Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD) [39]; ‘Critical Appraisal and Data Extraction for Systematic Reviews of Prediction Modelling Studies: The CHARMS Checklist’ [40], and ‘Protocol for a systematic review on the methodological and reporting quality of prediction model studies using machine learning techniques’ [41]. Table 3 details the data extraction categories. Primary data extraction was conducted by DN this was then validated by LW.

Table 3
Data Extraction summary

Category	*Description/example*
Title	Title of journal/conference entry.
Journal/ Conference	Publisher.
Outcome Benchmark for depression	How outcome was measured (e.g., PHQ-9, ICD code, HADS)
Demographic	Characteristics of the participant pool including age, gender, ethnicity etc. where specified.
Predictors	Types of predictors used by models and identification of any groupings or subsets they might fall into.
Study Design	Case/Control, Case Series, Cohort etc.
Data Source	EHRs, EMRs, Clinical Notes, Clinical Records
Sample Size Train or Total	Number included in training/total dataset.
Sample Size Test/Validate	Number included in test/validation dataset
Missing Data	Explanation of how instances of missing data were addressed.
Model Development Pre-Process	Information relating to the methods used for pre-processing, preparing, cleaning, extracting data (e.g., natural language and text processing methods).
Model Development Analysis (Fitting)	Information relating to the statistical methods used, ML (statistical techniques and/or broader AI e.g., neural networks). If relevant additional data pre-processing/preparation.
Performance Metric	How model measured/reported (e.g., odds ratio, AUC ROC, Sensitivity, Specificity, Accuracy).
Baseline/Comparator	Criteria used to evaluate/compare model. How model assessed against outcome.
Validation	Information relating to the use of validation methods, independent testing and separate hold out sets.
Results	The results reported (may be in summary form).
Data Availability	Information relating to data availability, any repository/contact details and conditions that might apply.
Code Availability	Information relating to code availability, any repository/contact details and conditions that might apply.
Abstract	Text of study abstract.

Category	*Description/example*
Full Reference (and Citation)	Supporting unambiguous identification of paper and providing source for citations in tables/figures/text.

Quality of studies

The Oxford Centre for Evidence-Based Medicine (OCEBM) system [42] was used to assess quality (previously used for a systematic review about artificial intelligence and suicide prevention by [43] as many of the models were developed and evaluated in a clinical setting and so merit a level of formal assessment. This ranked the evidence on a scale of 1 to 5, lowest to highest. The results were added to the data extraction table.

Results

The search protocol together with numbers of studies identified, selected, assessed, included/excluded is presented in Fig. 1, compatible with PRISMA standard [44)].

Searches

A total of 744 research papers were identified in the first stage of the literature search (711 after duplicates were removed). Screening content of abstracts and, subsequently, main body of each article, reduced the sample to 18 eligible articles. The backwards citation search of the selected papers resulted in one additional paper (so, 19 in total).

Review articles are not included in the final total but were used for supporting research and were recorded.

Selected studies overview

This review summarised studies that use ML methods to train statistical models for predicting depression based on individual-level EHR data from primary care (11 studies) and from a combination of primary and secondary care (8 studies). Table 4 summarizes key features of each study. We now turn to a detailed overview of each of the components described in Table 4.

Table 5
Grouping of predictors from the studies.

Predictor group	*Commentary*
Comorbidities	Comorbidities were included in thirteen studies. They included long-term conditions, such as diabetes, asthma, epilepsy, and chronic pain were commonly used, often when the study authors highlighted theoretical links with depression.
Demographic	Demographic predictors were used in sixteen studies. On some occasions, specific demographic variables were excluded due to insufficient availability/coverage (often the case for ethnicity). Gender was included as a predictor and occasionally also as a means of creating gender-specific models (e.g., Nichols *et al.* [55]). Social deprivation was also used as a predictor, and information about missed immunization(s) was used in two studies, Nemesure *et al.* [54] and Nichols *et al.*[55], as a proxy for social deprivation. The age range of cases was often an integral part of the study's specific aims. Some studies specifically focussed on older patients. For instance, Sau and Bhakta [58] used data with an average age of 68.5 years (standard deviation 4.85 years), whereas Nichols *et al.* [55] focused on early diagnosis among young people, between 15 to 24 years of age. Some studies narrowed the analysis to a narrow age bracket, others included a wide range of ages. For example, Hochman *et al*. [48], who studied postpartum depression reported an average age of 29.4 years (standard deviation, 5.4) whereas Xu *et al.* [61] used data from participants whose age ranged from 18 to over 65.
Family History	Family history was used in five studies and included family history of abuse (physical/sexual) and drug/substance abuse, often because the study authors cited theoretical links with depression. This group of predictors was often under recorded, as reported in the Nichols *et al.* [54] study where family history data was removed from the model due to low prevalence (< 0.02%) in their data. Insufficient family history data was also highlighted as a limitation in other studies [49,51].
Obstetric specific	Obstetric specific were used in five studies focussed on the prediction of postpartum depression, and these included predictors such as premature birth, use of specific drugs during pregnancy and obesity. This type of predictor was also used in non-postpartum depression studies e.g., Abar *et al.* [45].
Other (e.g., blood pressure)	Other predictors were used in eleven studies and included, e.g., measurements of physical characteristics such as blood pressure, cholesterol, results of assays, and height/weight.
Individual psychiatric symptoms or other diagnoses	Psychiatric symptoms/diagnoses were used in fifteen studies. These include both depression related symptoms such as: anxiety, low mood, self-harm, sleeping and eating disorders, too little sleep etc. They also include the broader range of conditions including post-traumatic stress syndrome, obsessive compulsive disorder, personality disorders and psychoses. Within individual studies there may/may not be a distinction made between these two subgroups.

Predictor group	*Commentary*
Smoking	Smoking was used in seven studies. However, it was identified, for instance by Nichols *et al.* [54], that data may be incomplete for all participants and that this might impact the ability to reliably assess correlations with depression, to mitigate this they used “missing smoker” data as a separate predictor.
Social/family	Social and family related factors were used in seven studies these included bereavement, divorce, single parent, police or social services involvement and similar.
Somatic	Somatic conditions were used in fourteen studies these include physical conditions such as, abdominal pain, back pain, dyspepsia, eczema, headaches, and others.
Substance/alcohol abuse	Alcohol/substance abuse was used in seven studies, participants identified as having drug/alcohol abuse problems. Typically categorical, but some studies included levels of abuse and/or combinations of the two.
Visit frequency	Visit frequency was used in six studies and shown to be a significant contributor to model performance. This is an integer variable based on number of visits in a specified period to the care facility (e.g., NHS GP).
Word list/text	Word list/text derived data was used in only one study, Geraci *et al.* [46], this was a source of data that was then analysed, using natural language processing, to extract predictors from clinical notes. It is based on language/defined terms specific.

Note: There may be overlap or gaps in these groupings as the predictors used and the reason for their use is study specific and not always explained.

Depression Definition

The definition of depression and the method of its classification varied across the studies in this review. A combination of depression diagnosis definitions based on NHS Read codes [64], SNOMED (Systematized Nomenclature For Medicine) [27] codes, ICD [10] or DSM [11] based assessments and/or the prescription of antidepressants (ADs) was used in 16 of the 19 studies. Only one study, by Xu *et al.* [61], used antidepressant prescription alone as a case definition. Three other studies relied on the use of a validated questionnaire such as the PHQ-9 [65] or HADS [14].

Predictors

Here we report on aspects of the predictors including their definition, how we grouped them and their frequency of use.

Definitions

Most predictors were derived from a combination of variables present in the EHR databases (e.g., SNOMED/NHS Read codes and/or prescription of a drug in a similar way to the definition used for depression) and were typically categorical. In some cases, additional parameters specifying a time frame for the predictor were also available. Some predictors were defined by identifying components by pre-processing clinical notes/other textual information. A few studies used non categorical predictors such as physiological measurements for example Body Mass Index (BMI), blood pressure, and cholesterol as predictors. This was usually where participants were receiving some form of secondary care, such as in pregnancy for PPD prediction.

Groups

No formal method for grouping was evident in the studies and, due to the large number of diverse predictors used in different papers, for clarity these were organised into the following groups. Specifically: comorbidity, demographic, family history, other (e.g., blood pressure), psychiatric, smoking, social/family, somatic, obstetric specific, substance/alcohol abuse, visit frequency and word list/text. Due to this flexibility in definition, there are overlaps between studies concerning which category a predictor might fall, for example a blood test may be in “other, or “obstetric specific”. Table 5 shows the predictors groups and commentary on their content.

Data

The studies in this review used data sets from EHRs systems, insurance claims databases and health service (primary and secondary) providers. As such they store, organise, and define data in a variety of ways that are not expected to be consistent with each other. Most of this data is categorical in nature, though some predictors such as blood pressure, are usually continuous variables within a range. It is noted that the individual EHRs systems are proprietary in nature and there is no universally accepted extant standard detailing how data should be categorised, stored, and organised for them. There are organisations developing, promoting, and gaining accreditation, for example Health Level Seven International [(66)] with ANSI (American National Standards Institute) [67]. However, none of these are globally adopted and the World Health Organization standard that did exist, E1384, was withdrawn in 2017 [68]. Lack of standardisation is currently a barrier to portability for individual applications.

Missing or erroneous data

Missing data either related to missing patients and/or missing predictor data. Nemesure *et al.* [54] estimated that, for their data set, missing values were present in 5% of the data overall and for 20 out of the 59 predictors they used. In some studies, missing data led to exclusion of cases from the analysis. For example, Koning *et al*. [51] excluded patients whose records did not identify gender or had no postcode registered. Huang *et al.* [47] removed entries where patients had less than 1.5 years of visit history. Wang *et al.* [60] excluded from the analysis PPD patients for whom there was no third trimester data.

In Nichols *et al.* [55]. missing smoking status was used to infer non-smoking on the basis this was less likely to be missed for smokers/those with smoking related disorders. Missing data also led to exclusion of predictors. Again, in Nichols *et al*. [55], the authors did not use ethnicity as it was missing in over 63% of patients. Similarly, Zhang *et al.* [63] excluded ethnicity from their USA dataset for the same reasons.

Many studies (e.g., Koning *et al.* [69] , Meng *et al*. [53], Nichols *et al.* [55]) raised concerns that errors in predictor data could affect performance, generalizability, and reliability of the models. Errors and missing data were identified as due to misclassification, measurement errors, data entry and bias; all of which can be difficult identify and/or correct in EHR data [70]. Other studies varied in the strategies used for dealing with missing data. Common approaches were to estimate the level for a missing point or simply acknowledge that remedial action was not available. Nemesure *et al.* [54] used an imputation approach for-their numerical data, such as blood pressure. Where remedial action is not possible then the patient can be excluded from the study, e.g. Hochman *et al*. [48].

Sources of bias

Some of the studies, for instance, Huang *et al.* [47] and Koning *et al.* [51] raised the question about data bias due to collection processes, such as diagnosis, data interpretation and system input. Other studies recognised sources of bias impacting accuracy and generalizability. Jin *et al*. [49] identified that as the population in their study were mainly Hispanic and there was incompleteness of comorbidity predictor data (e.g., for diabetes), both performance and generalizability would be affected. Zhang *et al.* [63] acknowledged that sourcing their data from an urban academic medical centre could introduce result in a limited generalizability of their findings. Hochman *et al.* [48] suggested that their use of an exclusion criteria removing severely depressed patients based on the prescription of specific drugs could also create bias. Zhang *et al.* [62] chose to exclude ethnicity from their models due to coding inconsistencies and errors; making a bias in that area a potential issue. Huang *et al.* [47], defined depression based solely on antidepressant usage and suggested their sample would be skewed towards the more severely depressed because the sample excluded those whose condition was treated with only psychotherapy or those without any treatment. A similar concern regarding changing definitions for the detection of depression during their study period was expressed by Xu *et al.* [61]. At a broader level, 20 of the studies were from "WEIRD" (Western, Educated, Industrialised, Rich, Democratic) countries with the majority (15) from the USA. The remainder were from countries with highly developed IT and healthcare industries such as Brazil, Israel, and India.

Data sharing

The nature of the data, data protection and requirements for anonymity, and privacy issues limited access to source data though details of sources themselves were more often made available (e.g., Hochman *et al*. [48], Nichols *et al*. [55]).

Modelling

In this review, we identified a wide array of statistical techniques used on EHR data (see table 4). Many different types of supervised ML were used for classification of depression versus control, including regression models (13 studies) and Random Forest (8 studies), XGBoost (8 studies) and SVM (7 studies) were the most common techniques. Use of multiple techniques in a single paper was also common, for instance Xu *et al.* [61] and Zhang *et al*. [62] used four or more methods. Geraci *et al*. [46] was the only study to use a deep neural network-based deep learning approach as the primary component of their model. Figure 3 summarises methods used in the selected studies.

Temporal sequence was referred to in two studies [45,56] though other studies refer to time between predictors and diagnosis (e.g., Meng *et al*. [52]). In other studies patterns of predictors were used to determine their predictive probabilities of depression, sometimes using time constraints, such as a primary care visit "within the last twelve months" or specifically including time distant events such as birth trauma (*Koning et al*. [51], Nichols *et al.* [54]). Only one study, Półchłopek *et al.* [56], considered temporal sequence in EHRs. Though Abar *et al*. [45] speculated that temporal sequence might be used to improve performance by taking causal sequence into consideration.

Most studies (17 out of 19) validated their models, most commonly (12) by splitting data into a training and a testing set. Cross validation data sets for model testing were also used (11 out of 19). Generally testing and validation was carried out by the same team as created the models, only Sau and Bhakta [58] had diagnostic accuracy checked by an independent team. Only one study used a separate data set for testing rather than splitting the original data set (Zhang *et al*. [63].

Code sharing

Code was made available by the majority (12) of studies. In some cases, just the details of the packages that implemented the ML algorithm were provided. For example, Jin *et al.* [62] reference the R package MASS, rather than the providing the complete code.

Performance

Several performance metrics was used to evaluate ML models of depression. Among those, researchers reported confusion matrices; area under the curve – receiver operating characteristics (AUC-ROC); and Odds Ratios/Variable Importance for predictors.

Confusion Matrix derived metrics (True Positives, True Negatives, False Positives and False Negatives) were used in sixteen of the studies, sometimes in conjunction with other measures particularly AUC-ROC. Many performance metrics are derived from this information, including accuracy, F1, sensitivity, specificity, and precision. Sensitivity and specificity were commonly reported, possibly because they give information relating to the discriminative performance of the model and are well understood by practitioners [71]).

For sensitivity, reported values range from 0.35 Hochmam *et al.* [48] to 0.94 Geraci *et al*. [46]. For specificity, reported values range from 0.39 Wang *et al.* [60] to 0.91 Hochman *et al.* [48]. Sensitivity was usually higher than specificity across the models with the exceptions being: Hochman *et al.* [48)] who reported a high specificity figure of 0.91 with a low sensitivity of 0.35 using a gradient boosted decision tree algorithm; and Nemesure *et al*. [54] reported specificity of 0.7 and sensitivity of 0.55. The highest accuracy at 0.91 was reported by Sau and Bhakta [58] and the lowest was 0.56 (Zhang *et al.* [63]). This metric only gives a broad overall picture of correctly predicted results vs. all predictions made and gives no indication of the more useful true/false positive rates; it was presented in only six studies.

For the studies that reported performance in terms of AUC- ROC metric (14) the low extreme for any model was 0.55, specifically from a benchmark model predicting depression in the 12-15 years age group (Półchłopek *et al.* [56]). The highest AUC-ROC score was 0.94 (Zhang *et al.* [63], Kasthurirathne *et al.* [50])The overall range AUC-ROC values reported was 0.70 to 0.90. The average AUC-ROC value was 0.78 with a standard deviation of 0.07. Figure 4 shows the average AUC values achieved in each study.

Generalizability and Interpretability

Generalizability was mentioned in several studies, for example Jin *et al.* [49] and Zhang *et al.* [63]. The points already illustrated under, "sources of bias", for example, demographically specific participants, and, factors relating to missing data and granularity of data, such as only having social deprivation data at practice level have negative consequences for generalizability.

Interpretability was identified as a concern in several studies (e.g., Koning *et al.* ([51]), Nemesure *et al.* [54], Meng *et al.* [52]). For interpretability Nemesure *et al.* [54] used SHAP (Shapley Additive Explanations) scores which offers a decision chart for the model predictors [72]. None of the included studies provided visualisations other than AUC-ROC diagrams and bar charts, as such interpretability was not significantly addressed in the selected studies.

Quality of Studies

All the included studies achieved a score of 3 (11) or 4 (8) based on the OCEBM criteria (1 to 5 from highest to lowest) as far they could be applied to the selected studies, areas that related to diagnostic tests only (no interventions). This represents a moderate level of performance. Overall, the studies represented large sample sizes, usually case series or cohort trials and they applied a clinically recognised benchmark, had there been randomized trials studies could have been promoted to level 2.

Only 3 studies provided reference to the use of a formal assessment method such as TRIPOD [39]. suggesting that following standards is not yet widespread or that the frameworks are not yet sufficiently established or appropriate. This lack of consistent reporting is a limitation, and the use of standardised frameworks should become the expectation rather than the exception.

Discussion

In this review we have identified three areas: performance, generalizability, and interpretability as key components to consider for predictive models of depression built on the use of ML with EHR data. All three would need careful evaluation before moving from research to a clinical application environment.

Generalizability

To be widely deployed clinically, the models in the studies would need to be generalizable, i.e., be able to work reliably outside of their development environment. Kelly *et al.* [73] identified the ability to deal with new populations as one prerequisite for clinical success. Areas identified in the studies that could impact generalizability included demographics, sources of bias, inclusion/exclusion criteria, missing/incomplete data, the definition of depression and predictors. All of these were identified in the included studies, for instance, Jin *et al.* [49] identified Hispanic participants being highly represented in their data and Zhang *et al.* [62] excluding ethnicity from their models.

As noted in the Performance sub-section of the Results, the ML method itself did not seem to be overly critical for outcome performance using the EHR data sets in the included studies and it is provisionally suggested that the method itself may be more generalizable than the data to which it is fitted.

Another area that can limit generalizability is the wide variety of EHR data. This varies depending on source for example insurance derived, a state health service such as the NHS, or a proprietary standard such as SNOMED etc. The coding may, or may not, incorporate a recognised medical standard such as the ICD [10] or DSM [11] amongst others that can be found in the included studies. Although not derived from the studies directly, it was noted that there is no global standard in current use in place covering content/structure/format for EHR data. Consequently, it is likely that models are data source specific to a greater or lesser extent. Further work needs to consider how this can be addressed.

The studies in this review differed in how depression was defined and by the range of predictors selected and their definitions. As mentioned, a commonly used approach was to use a combination of EHR data entry codes covering diagnoses in combination with prescription of an antidepressant. This can result in too many cases as being diagnosed as depressed due to antidepressants being used for a wider range of conditions. Similar issues apply for the definition of predictors. In combination this restricts the generalizability of any models produced.

Another factor for generalization is the robustness of the models and their replicability. None of the studies included replication of their results, only Sau and Bhakta [58] used an independent team for the verification of results, though the majority employed recognised validation techniques. Reducing bias and independent validation should be a recommendation for future work involving the prediction of depression using ML with EHRs.

Interpretability

Interpretability was only identified as a concern in a few studies. However, clinical practitioners may wish to know the explanation for ML algorithm’s predicted diagnosis so they can fit it into a broader diagnostic picture rather than treating it as a “black box” [74]. Similarly, Vellido [75] and Stiglic *et al.* [76] also considered that interpretability and visualisation are important for effective implementation of medical ML applications. This may be as simple as listing the specific predictors that contributed to the outcome, for example, anxiety, low mood, chronic pain or similar. Of the included studies Nemesure *et al.* [54] used SHAP (Shapley Additive Explanations) scores which have been used in clinical applications [77] to aid interpretability, again by identifying the most important predictors. However, none of the other included studies provided visualisations other than AUC-ROC diagrams and bar charts of predictors. It is recommended that future studies should be made that not only develop predictive models but also include trialling their use, for example with primary practitioners, support staff and/or patients, offering different forms of interpretable/black box output and assessing acceptability.

Performance

A limiting factor on performance in the included studies, relates to the definition of depression itself and the predictors used. Defining depression accurately is critical as this definition is used to train the ML application, a point raised by Meng *et al.* [52]. In the studies reviewed here, typically a combination of diagnostic and drug codes within the EHRs were used. Using prescription of antidepressants as part of the definition may misidentify too many cases. ADs are prescribed for other conditions including anxiety [78,79], chronic pain [80,81], obsessive compulsive disorder [82,83], post-traumatic stress disorder [84,85] and inflammatory bowel disease [86]. Of the included papers Xu *et al.* [61] suggested that under-identification of depression cases could also occur for patients receiving treatment via private care or an alternate service provider.

The prevalence of predictors can be artificially boosted, as suggested by Koning *et al*. [51] and Nichols *et al*. [55] where primary care physicians who think a patient has depression may identify or suspect a precursor or comorbidity, for example, with other mental health conditions like low mood or anxiety. There is strong evidence that family history of depression, alcohol, drug, physical and sexual abuse, and co-morbidity with other mental health conditions, are strong predictors of depression [87–90]. However, this data appears to be under recorded resulting in removal of important predictors due to low prevalence - again in Nichols *et al.* [55] removed family history data due to its low prevalence (< 0.02%). This would be expected to have a negative impact on performance. Identifying consistent and valid definitions for depression and any predictors used is a necessity.

The studies in this review reported an overall model performance where AUC-ROC value was 0.78 with a standard deviation of 0.07 (figure 2). This compares well with primary care where up to half of depression cases are missed at baseline consultation, improving to around two thirds being diagnosed at follow up [35,37]. An earlier paper [91] reported that only 39.1% of cases of ICD10 current depression were identified by primary care practitioners. Based on the studies we identified potential areas that might support improvements in the performance of the models.

Although some studies suggested that using more sophisticated techniques should improve performance, we noted that simpler methods such as logistic regression were often comparable to those obtained using more complex ones such as Random Forest and XG Boost (e.g., *Zhang et al*. [63]. Christodoulou *et al.* [92] echoed this conclusion in their systematic review of clinical prediction using ML where they saw similar performance for logistic regression compared with ML models such as, artificial neural networks, decision trees, Random Forest, and support vector machines (SVM). Geraci *et al.* [46] employed a deep neural network (deep learning) as their main modelling technique and Nemesure *et al*. [54] used it as a component in a larger ensemble model.   However, neither demonstrated performance benefits from its use. Even if higher performance could be obtained using deep learning it is important to note that small amounts of noise or small errors in the data can cause significant reliability issues due to misclassification due to very small perturbations in the data [94,95]. The use of more sophisticated techniques to improve performance is not supported by this review.

How else might performance be improved? The use of non-anonymised data, sourced from within a primary or secondary care facility, something that is more achievable in a clinical than a research setting, could be beneficial. For example, in the Nichols *et al.* [55] study social deprivation indices were only available at a regional/practice level and inspection of their model suggests that social deprivation has little impact on prediction of depression. This is inconsistent with expectation, as supported by Ridley *et al.* [95] who showed that there is a link between increased social deprivation and the probability of developing depression. Having this data at an individual level might be expected to increase the performance of a model. However, this is likely to only be achievable in a clinical trial of an application. Alternatively, the use of synthetically generated EHR data [96,97] removes the patient confidentiality and related ethical constraints that come with real data and would allow all aspects of a model to be fully evaluated as if with non-anonymous patient data.

Another approach is using more information relating to time in predictive models; EHRs typically time stamp entries so it is known when a predictor is activated. Półchłopek *et al*. [56], considered temporal sequence in EHRs. They were concerned that techniques including support vector machines and random forest identify predictors that affect the outcome but do not identify the effect of sequence on that outcome. They looked at the improvement that could be found by using temporal patterns in addition to non-time specific predictors and noted a small positive effect. Abar *et al.* [45] also speculated that temporal sequence might be used to improve model performance. There are techniques that might be used to do this. For example, time series analysis methods such as Gaussian processes, which are capable of coping with the sparse nature of EHR data [98] have been used to make predictions for patients with heart conditions. We recommend exploring the use of more time dependent factors in building predictive ML models for depression.

Although missing data is more of a concern in terms of generalizability, some studies identified it as an opportunity to improve performance. Kasthurirathne *et al.* [99] noted that missing EHR data can reduce model performance and suggested that this could be mitigated by merging with other data sources, for example, related insurance claims. Nichols *et al.* [55] used missing smoking data as a predictor and it had a positive effect in their model. Missing data is potentially of significance of itself and is an opportunity for further study.

Strengths and Limitations

As far as we are aware this is the first systematic review focussed on the use of EHRs to predict depression using ML methods. The choice of journal databases and the date range covered by the searches means that the studies identified provide a sound basis for comparison. The data extraction protocol was informed by established standards [39–41] to best identify data needed to support meaningful and repeatable analyses.

A limitation of this study is that inclusion criteria focused on study titles and key words which may have led to some ML studies using EHRs being missed. This was mitigated using backwards and forwards citation searches. Additionally, the variety of study designs including case control, cohort, and longitudinal studies precluded the possibility of using some of the more traditional quality assessment tools; we did however, as stated in methods, use OCEBM which has been used in previous ML systematic reviews. The categorization, definition, and identification of the numbers of predictors used within models was sometimes difficult to establish, leading to limitation in the scope of this information presented. It is also likely that the included studies are culturally specific as they focused on “WEIRD” populations.

Conclusion

In conducting this systematic review, we have shown that there is a body of work that supports the potential use of ML techniques with EHRs for the prediction of depression. This approach can deliver performance that is comparable to, or better than that found in primary care. It is clear there is scope for improvement both in terms of adoption of standards for both conducting and reporting the research and the data itself. This would involve greater promotion, and development, of standards for research such as TRIPOD [39] and, for data interchange, Health Level Seven International [66], and their further development to support ML/EHR applications. Future work could pay more attention to generalizability and interpretability, both of which need to be addressed for successful implementation in the clinic.   It is also worth investigating areas where performance can be improved, for example by including temporal sequence within the models, better selection of predictors and the use of non-anonymised/synthetic data. Our review suggests depression prediction using ML/EHRs is a worthwhile area for future development.

Declarations

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Availability of data and materials

All data generated or analysed during this study are included in this published article [and its supplementary information files].

Competing interests

The authors declare that they have no competing interests.

Funding

This work was supported by the Engineering and Physical Sciences Research Council (EPSRC) Doctoral Training Partnership Award (#2300953). Funding was awarded to CT as PI.

The funding source had no role in study design; the collection, analysis, and interpretation of data; the writing of the report; or the decision to submit the article for publication.

Authors' contributions

DN and CT defined the systematic review scope and designed the methods. DN managed the literature searches and analyses. DN and LW undertook the statistical analysis, and DN wrote the first draft of the manuscript. CT, CM and LW reviewed and proofread subsequent versions of the manuscript prior to submission.

All authors contributed to and have approved the final manuscript.

Acknowledgements

University of Warwick provided the library, information technology and office facilities that supported the contributors in the production of this study.

References

1. Lim GY, Tam WW, Lu Y, Ho CS, Zhang MW, Ho RC. Prevalence of Depression in the Community from 30 Countries between 1994 and 2014. Sci Rep. 2018 Feb 12;8(1):2861.
2. Vigo D, Thornicroft G, Atun R. Estimating the true global burden of mental illness. Lancet Psychiatry. 2016 Feb 1;3(2):171–8.
3. Ferrari AJ, Charlson FJ, Norman RE, Patten SB, Freedman G, Murray CJL, et al. Burden of Depressive Disorders by Country, Sex, Age, and Year: Findings from the Global Burden of Disease Study 2010. PLOS Med. 2013 Nov 5;10(11):e1001547.
4. Chesney E, Goodwin GM, Fazel S. Risks of all-cause and suicide mortality in mental disorders: a meta-review. World Psychiatry. 2014;13(2):153–60.
5. Organization WH. Depression and other common mental disorders: global health estimates. 2017 [cited 2022 Nov 11]; Available from: https://policycommons.net/artifacts/546082/depression-and-other-common-mental-disorders/1523689/
6. McCrone P, Dhanasiri S, Patel A, Knapp M, Lawton-Smith S. Paying the price: the cost of mental health care in England to 2026 [Internet]. The King’s Fund; 2008 [cited 2021 Nov 29]. Available from:

https://kclpure.kcl.ac.uk/portal/en/publications/paying-the-price-the-cost-of-mental-health-care-in-england-to-2026(ebb0265b-c5be-4326-96f4-21d4f9ed4744).html

7. Fineberg NA, Haddad PM, Carpenter L, Gannon B, Sharpe R, Young AH, et al. The size, burden and cost of disorders of the brain in the UK. J Psychopharmacol (Oxf). 2013 Sep 1;27(9):761–70.
8. McGorry PD, Hickie IB, Yung AR, Pantelis C, Jackson HJ. Clinical staging of psychiatric disorders: a heuristic framework for choosing earlier, safer and more effective interventions. Aust N Z J Psychiatry. 2006;40(8):616–22.
9. McGorry PD. Early Intervention in Psychosis. J Nerv Ment Dis. 2015 May;203(5):310–8.
10. International Classification of Diseases (ICD) [Internet]. [cited 2023 Jan 20]. Available from: https://www.who.int/standards/classifications/classification-of-diseases
11. DSM [Internet]. [cited 2023 Jan 20]. Available from: https://www.psychiatry.org:443/psychiatrists/practice/dsm
12. Andrews G, Peters L, Guzman AM, Bird K. A comparison of two structured diagnostic interviews: CIDI and SCAN. Aust N Z J Psychiatry. 1995 Jan 1;29(1):124–32.
13. Robins LN, Wing J, Wittchen HU, Helzer JE, Babor TF, Burke J, et al. The Composite International Diagnostic Interview: An Epidemiologic Instrument Suitable for Use in Conjunction With Different Diagnostic Systems and in Different Cultures. Arch Gen Psychiatry. 1988 Dec 1;45(12):1069–77.
14. Zigmond AS, Snaith RP. The Hospital Anxiety and Depression Scale. Acta Psychiatr Scand. 1983;67(6):361–70.
15. Smarr KL, Keefer AL. Measures of depression and depressive symptoms: Beck Depression Inventory-II (BDI-II), Center for Epidemiologic Studies Depression Scale (CES-D), Geriatric Depression Scale (GDS), Hospital Anxiety and Depression Scale (HADS), and Patient Health Questionnaire-9 (PHQ-9). Arthritis Care Res. 2011 Nov;63 Suppl 11:S454-466.
16. BECK AT, WARD CH, MENDELSON M, MOCK J, ERBAUGH J. An Inventory for Measuring Depression. Arch Gen Psychiatry. 1961 Jun 1;4(6):561–71.
17. Spitzer RL, Kroenke K, Williams JBW, and the Patient Health Questionnaire Primary Care Study Group. Validation and Utility of a Self-report Version of PRIME-MDThe PHQ Primary Care Study. JAMA. 1999 Nov 10;282(18):1737–44.
18. Kroenke K. PHQ-9: global uptake of a depression scale. World Psychiatry. 2021 Feb;20(1):135–6.
19. Kocalevent RD, Hinz A, Brähler E. Standardization of the depression screener Patient Health Questionnaire (PHQ-9) in the general population. Gen Hosp Psychiatry. 2013 Sep 1;35(5):551–5.
20. Arroll B, Goodyear-Smith F, Crengle S, Gunn J, Kerse N, Fishman T, et al. Validation of PHQ-2 and PHQ-9 to Screen for Major Depression in the Primary Care Population. Ann Fam Med. 2010 Jul 1;8(4):348–53.
21. Levis B, Benedetti A, Thombs BD. Accuracy of Patient Health Questionnaire-9 (PHQ-9) for screening to detect major depression: individual participant data meta-analysis. BMJ. 2019 Apr 9;365:l1476.

22. Bohlmeijer ET, Fledderus M, Rokx TAJJ, Pieterse ME. Efficacy of an early intervention based on acceptance and commitment therapy for adults with depressive symptomatology: Evaluation in a randomized controlled trial. Behav Res Ther. 2011 Jan 1;49(1):62–7.

23. Davey CG, McGorry PD. Early intervention for depression in young people: a blind spot in mental health care. Lancet Psychiatry. 2019 Mar 1;6(3):267–72.

24. McGorry P, van Os J. Redeeming diagnosis in psychiatry: timing versus specificity. The Lancet. 2013 Jan 26;381(9863):343–5.

25. Office-based Physician Electronic Health Record Adoption | HealthIT.gov [Internet]. [cited 2021 Oct 27]. Available from: https://www.healthit.gov/data/quickstats/office-based-physician-electronic-health-record-adoption

26. Jha AK, Doolan D, Grandt D, Scott T, Bates DW. The use of health information technology in seven nations. Int J Med Inf. 2008 Dec;77(12):848–54.

27. SNOMED Home page [Internet]. SNOMED. [cited 2021 Nov 2]. Available from: https://www.snomed.org/

28. Kruse CS, Stein A, Thomas H, Kaur H. The use of Electronic Health Records to Support Population Health: A Systematic Review of the Literature. J Med Syst. 2018 Sep 29;42(11):214.

29. QRISK3 [Internet]. [cited 2021 Oct 27]. Available from: https://qrisk.org/three/index.php

30. Pike MM, Decker PA, Larson NB, St. Sauver JL, Takahashi PY, Roger VL, et al. Improvement in Cardiovascular Risk Prediction with Electronic Health Records. J Cardiovasc Transl Res. 2016 Jun 1;9(3):214–22.

31. Klompas M, Eggleston E, McVetta J, Lazarus R, Li L, Platt R. Automated Detection and Classification of Type 1 Versus Type 2 Diabetes Using Electronic Health Record Data. Diabetes Care. 2013 Apr 1;36(4):914–21.

32. Hippisley-Cox J, Coupland C, Vinogradova Y, Robson J, Minhas R, Sheikh A, et al. Predicting cardiovascular risk in England and Wales: prospective derivation and validation of QRISK2. BMJ. 2008 Jun 28;336(7659):1475–82.

33. Shatte ABR, Hutchinson DM, Teague SJ. Machine learning in mental health: a scoping review of methods and applications. Psychol Med. 2019;49(9):1426–48.

34. Cho G, Yim J, Choi Y, Ko J, Lee SH. Review of Machine Learning Algorithms for Diagnosing Mental Illness. Psychiatry Investig. 2019 Apr;16(4):262–9.

35. Kessler D, Bennewith O, Lewis G, Sharp D. Detection of depression and anxiety in primary care: follow up study. BMJ. 2002 Nov 2;325(7371):1016–7.

36. Kessler RC, Bromet EJ. The epidemiology of depression across cultures. Annu Rev Public Health. 2013;34:119–38.

37. Mitchell AJ, Rao S, Vaze A. Can general practitioners identify people with distress and mild depression? A meta-analysis of clinical accuracy. J Affect Disord. 2011 Apr 1;130(1):26–36.

38. Booth A, Clarke M, Dooley G, Ghersi D, Moher D, Petticrew M, et al. The nuts and bolts of PROSPERO: an international prospective register of systematic reviews. Syst Rev. 2012 Feb 9;1(1):2.

39. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. BMJ [Internet]. 2015 [cited 2021 Apr 26];350. Available from: https://www.jstor.org/stable/26517836

40. Moons KGM, Groot JAH de, 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 Oct 14;11(10):e1001744.

41. Navarro CLA, Damen JAAG, Takada T, Nijman SWJ, Dhiman P, Ma J, et al. Protocol for a systematic review on the methodological and reporting quality of prediction model studies using machine learning techniques. BMJ Open. 2020 Nov 1;10(11):e038832.

42. OCEBM Levels of Evidence — Centre for Evidence-Based Medicine (CEBM), University of Oxford [Internet]. [cited 2022 Nov 17]. Available from: https://www.cebm.ox.ac.uk/resources/levels-of-evidence/ocebm-levels-of-evidence

43. Bernert RA, Hilberg AM, Melia R, Kim JP, Shah NH, Abnousi F. Artificial Intelligence and Suicide Prevention: A Systematic Review of Machine Learning Investigations. Int J Environ Res Public Health. 2020 Jan;17(16):5929.

44. Moher D, Liberati A, Tetzlaff J, Altman DG. Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. Int J Surg. 2010 Jan 1;8(5):336–41.

45. Abar O, Charnigo RJ, Rayapati A, Kavuluru R. On Interestingness Measures for Mining Statistically Significant and Novel Clinical Associations from EMRs. In: Proceedings of the 7th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics [Internet]. New York, NY, USA: Association for Computing Machinery; 2016 [cited 2021 Jul 14]. p. 587–94. (BCB ’16). Available from: https://doi.org/10.1145/2975167.2985843

46. Geraci J, Wilansky P, de Luca V, Roy A, Kennedy JL, Strauss J. Applying deep neural networks to unstructured text notes in electronic medical records for phenotyping youth depression. Evid Based Ment Health. 2017;20(3):83–7.

47. Huang SH, LePendu P, Iyer SV, Tai-Seale M, Carrell D, Shah NH. Toward personalizing treatment for depression: predicting diagnosis and severity. J Am Med Inform Assoc. 2014 Nov 1;21(6):1069–75.

48. Hochman E, Feldman B, Weizman A, Krivoy A, Gur S, Barzilay E, et al. Development and validation of a machine learning-based postpartum depression prediction model: A nationwide cohort study. Depress Anxiety. 2021;38(4):400–11.

49. Jin H, Wu S, Vidyanti I, Di Capua P, Wu B. Predicting Depression among Patients with Diabetes Using Longitudinal Data. A Multilevel Regression Model. Methods Inf Med. 2015;54(6):553–9.

50. Kasthurirathne SN, Biondich PG, Grannis SJ, Purkayastha S, Vest JR, Jones JF. Identification of Patients in Need of Advanced Care for Depression Using Data Extracted From a Statewide Health Information Exchange: A Machine Learning Approach. J Med Internet Res. 2019;21(7):e13809.

51. Koning NR, Büchner FL, Vermeiren RRJM, Crone MR, Numans ME. Identification of children at risk for mental health problems in primary care—Development of a prediction model with routine health care data. EClinicalMedicine. 2019 Oct 1;15:89–97.

52. Meng Y, Speier W, Ong MK, Arnold CW. Bidirectional Representation Learning from Transformers using Multimodal Electronic Health Record Data to Predict Depression. ArXiv200912656 Cs [Internet]. 2020 Oct 30 [cited 2021 Jan 7]; Available from: http://arxiv.org/abs/2009.12656

53. Meng Y, Speier W, Ong M, Arnold CW. HCET: Hierarchical Clinical Embedding With Topic Modeling on Electronic Health Records for Predicting Future Depression. IEEE J Biomed Health Inform. 2021 Apr;25(4):1265–72.

54. Nemesure MD, Heinz MV, Huang R, Jacobson NC. Predictive modeling of depression and anxiety using electronic health records and a novel machine learning approach with artificial intelligence. Sci Rep. 2021 Jan 21;11(1):1980.

55. Nichols L, Ryan R, Connor C, Birchwood M, Marshall T. Derivation of a prediction model for a diagnosis of depression in young adults: a matched case–control study using electronic primary care records. Early Interv Psychiatry. 2018;12(3):444–55.

56. Półchłopek O, Koning NR, Büchner FL, Crone MR, Numans ME, Hoogendoorn M. Quantitative and temporal approach to utilising electronic medical records from general practices in mental health prediction. Comput Biol Med. 2020 Oct 1;125:103973.

57. Qiu R, Kodali V, Homer M, Heath A, Wu Z, Jia Y. Predictive Modeling of Depression with a Large Claim Dataset. In: 2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). 2019. p. 1589–95.

58. Sau A, Bhakta I. Predicting anxiety and depression in elderly patients using machine learning technology. Healthc Technol Lett. 2017;4(6):238–43.

59. Souza Filho EM de, Veiga Rey HC, Frajtag RM, Arrowsmith Cook DM, Dalbonio de Carvalho LN, Pinho Ribeiro AL, et al. Can machine learning be useful as a screening tool for depression in primary care? J Psychiatr Res. 2021 Jan 1;132:1–6.

60. Wang S, Pathak J, Zhang Y. Using Electronic Health Records and Machine Learning to Predict Postpartum Depression. Stud Health Technol Inform. 2019 Aug 21;264:888–92.

61. Xu Z, Wang F, Adekkanattu P, Bose B, Vekaria V, Brandt P, et al. Subphenotyping depression using machine learning and electronic health records. Learn Health Syst. 2020;4(4):e10241.

62. Zhang J, Xiong H, Huang Y, Wu H, Leach K, Barnes LE. M-SEQ: Early detection of anxiety and depression via temporal orders of diagnoses in electronic health data. In: 2015 IEEE International Conference on Big Data (Big Data). 2015. p. 2569–77.

63. Zhang Y, Wang S, Hermann A, Joly R, Pathak J. Development and validation of a machine learning algorithm for predicting the risk of postpartum depression among pregnant women. J Affect Disord. 2021 Jan 15;279:1–8.

64. SCIMP Guide to Read Codes | Primary Care Informatics [Internet]. [cited 2021 Nov 12]. Available from: https://www.scimp.scot.nhs.uk/better-information/clinical-coding/scimp-guide-to-read-codes

65. Kroenke K, Spitzer RL, Williams JBW. The PHQ-9. J Gen Intern Med. 2001;16(9):606–13.

66. Health Level Seven International - Homepage | HL7 International [Internet]. [cited 2022 Nov 17]. Available from: http://www.hl7.org/index.cfm

67. American National Standards Institute - ANSI Home [Internet]. [cited 2022 Nov 17]. Available from: https://www.ansi.org/

68. Standard Practice for Content and Structure of the Electronic Health Record (EHR) (Withdrawn 2017) [Internet]. [cited 2022 Nov 17]. Available from: https://www.astm.org/e1384-07r13.html

69. Koning NR, Büchner FL, Leeuwenburgh NA, Paijmans IJ, Dijk DA van D van, Vermeiren RR, et al. Identification of child mental health problems by combining electronic health record information from different primary healthcare professionals: a population-based cohort study. BMJ Open. 2022 Jan 1;12(1):e049151.

70. Wu H, Yamal JM, Yaseen A, Maroufy V. Statistics and Machine Learning Methods for EHR Data: From Data Extraction to Data Analytics. CRC Press; 2020. 329 p.

71. Harris M, Taylor G. Medical Statistics Made Easy: 3rd Edition [Internet]. Scion Publications; 2014 [cited 2023 Jan 20]. Available from: http://www.scionpublishing.com

72. Merrick L, Taly A. The Explanation Game: Explaining Machine Learning Models Using Shapley Values. In: Holzinger A, Kieseberg P, Tjoa AM, Weippl E, editors. Machine Learning and Knowledge Extraction. Cham: Springer International Publishing; 2020. p. 17–38. (Lecture Notes in Computer Science).

73. Kelly CJ, Karthikesalingam A, Suleyman M, Corrado G, King D. Key challenges for delivering clinical impact with artificial intelligence. BMC Med. 2019 Oct 29;17(1):195.

74. Cadario R, Longoni C, Morewedge CK. Understanding, explaining, and utilizing medical artificial intelligence. Nat Hum Behav. 2021 Dec;5(12):1636–42.

75. Vellido A. The importance of interpretability and visualization in machine learning for applications in medicine and health care. Neural Comput Appl. 2020 Dec 1;32(24):18069–83.

76. Stiglic G, Kocbek P, Fijacko N, Zitnik M, Verbert K, Cilar L. Interpretability of machine learning-based prediction models in healthcare. WIREs Data Min Knowl Discov. 2020;10(5):e1379.

77. Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, Nair B, et al. From local explanations to global understanding with explainable AI for trees. Nat Mach Intell. 2020 Jan;2(1):56–67.

78. Bandelow B, Michaelis S, Wedekind D. Treatment of anxiety disorders. Dialogues Clin Neurosci. 2017 Jun;19(2):93–107.

79. Ströhle A, Gensichen J, Domschke K. The Diagnosis and Treatment of Anxiety Disorders. Dtsch Ärztebl Int. 2018 Sep;115(37):611–20.

80. Sutherland AM, Nicholls J, Bao J, Clarke H. Overlaps in pharmacology for the treatment of chronic pain and mental health disorders. Prog Neuropsychopharmacol Biol Psychiatry. 2018 Dec 20;87:290–7.

81. Urits I, Peck J, Orhurhu MS, Wolf J, Patel R, Orhurhu V, et al. Off-label Antidepressant Use for Treatment and Management of Chronic Pain: Evolving Understanding and Comprehensive Review. Curr Pain Headache Rep. 2019 Jul 29;23(9):66.

82. Brakoulias V, Starcevic V, Albert U, Arumugham SS, Bailey BE, Belloch A, et al. Treatments used for obsessive–compulsive disorder—An international perspective. Hum Psychopharmacol Clin Exp. 2019;34(1):e2686.

83. Del Casale A, Sorice S, Padovano A, Simmaco M, Ferracuti S, Lamis DA, et al. Psychopharmacological Treatment of Obsessive-Compulsive Disorder (OCD). Curr Neuropharmacol. 2019 Aug 1;17(8):710–36.

84. Abdallah CG, Averill LA, Akiki TJ, Raza M, Averill CL, Gomaa H, et al. The Neurobiology and Pharmacotherapy of Posttraumatic Stress Disorder. Annu Rev Pharmacol Toxicol. 2019 Jan 1;59:171–89.

85. Ehret M. Treatment of posttraumatic stress disorder: Focus on pharmacotherapy. Ment Health Clin. 2019 Nov 1;9(6):373–82.

86. Jayasooriya N, Blackwell J, Saxena S, Bottle A, Petersen I, Creese H, et al. Antidepressant medication use in Inflammatory Bowel Disease: a nationally representative population-based study. Aliment Pharmacol Ther [Internet]. [cited 2022 Mar 15];n/a(n/a). Available from: https://onlinelibrary.wiley.com/doi/abs/10.1111/apt.16820

87. Milne BJ, Caspi A, Harrington H, Poulton R, Rutter M, Moffitt TE. Predictive Value of Family History on Severity of Illness: The Case for Depression, Anxiety, Alcohol Dependence, and Drug Dependence. Arch Gen Psychiatry. 2009 Jul 1;66(7):738–47.

88. van Dijk MT, Murphy E, Posner JE, Talati A, Weissman MM. Association of Multigenerational Family History of Depression With Lifetime Depressive and Other Psychiatric Disorders in Children: Results from the Adolescent Brain Cognitive Development (ABCD) Study. JAMA Psychiatry. 2021 Jul 1;78(7):778–87.

89. Weissman MM, Wickramaratne P, Gameroff MJ, Warner V, Pilowsky D, Kohad RG, et al. Offspring of Depressed Parents: 30 Years Later. Am J Psychiatry. 2016 Oct 1;173(10):1024–32.

90. Williamson DE, Ryan ND, Birmaher B, Dahl RE, Kaufman J, Rao U, et al. A Case-Control Family History Study of Depression in Adolescents. J Am Acad Child Adolesc Psychiatry. 1995 Dec 1;34(12):1596–607.

91. Sartorius N, Ustün TB, Organization WH. Mental illness in general health care : an international study [Internet]. Chichester: Wiley; 1995 [cited 2022 Feb 10]. Available from: https://apps.who.int/iris/handle/10665/36937

92. Christodoulou E, Ma J, Collins GS, Steyerberg EW, Verbakel JY, Van Calster B. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. J Clin Epidemiol. 2019 Jun 1;110:12–22.

93. Basu S, Pope P, Feizi S. Influence Functions in Deep Learning Are Fragile. ArXiv200614651 Cs Stat [Internet]. 2021 Feb 10 [cited 2022 Mar 28]; Available from: http://arxiv.org/abs/2006.14651

94. Ghorbani A, Abid A, Zou J. Interpretation of Neural Networks Is Fragile. Proc AAAI Conf Artif Intell. 2019 Jul 17;33(01):3681–8.
95. Ridley M, Rao G, Schilbach F, Patel V. Poverty, depression, and anxiety: Causal evidence and mechanisms. Science [Internet]. 2020 Dec 11 [cited 2020 Dec 16];370(6522). Available from: https://science.sciencemag.org/content/370/6522/eaay0214
96. Goncalves A, Ray P, Soper B, Stevens J, Coyle L, Sales AP. Generation and evaluation of synthetic patient data. BMC Med Res Methodol. 2020 May 7;20(1):108.
97. Walonoski J, Kramer M, Nichols J, Quina A, Moesel C, Hall D, et al. Synthea: An approach, method, and software mechanism for generating synthetic patients and the synthetic electronic health care record. J Am Med Inform Assoc. 2018 Mar 1;25(3):230–8.
98. Cheng LF, Dumitrascu B, Darnell G, Chivers C, Draugelis M, Li K, et al. Sparse multi-output Gaussian processes for online medical time series prediction. BMC Med Inform Decis Mak. 2020 Jul 8;20(1):152.
99. Kasthurirathne SN, Biondich PG, Grannis SJ, Purkayastha S, Vest JR, Jones JF. Identification of Patients in Need of Advanced Care for Depression Using Data Extracted From a Statewide Health Information Exchange: A Machine Learning Approach. J Med Internet Res. 2019;21(7):e13809.

Table 4

Table 4 is available in Supplementary Files section.

Figures

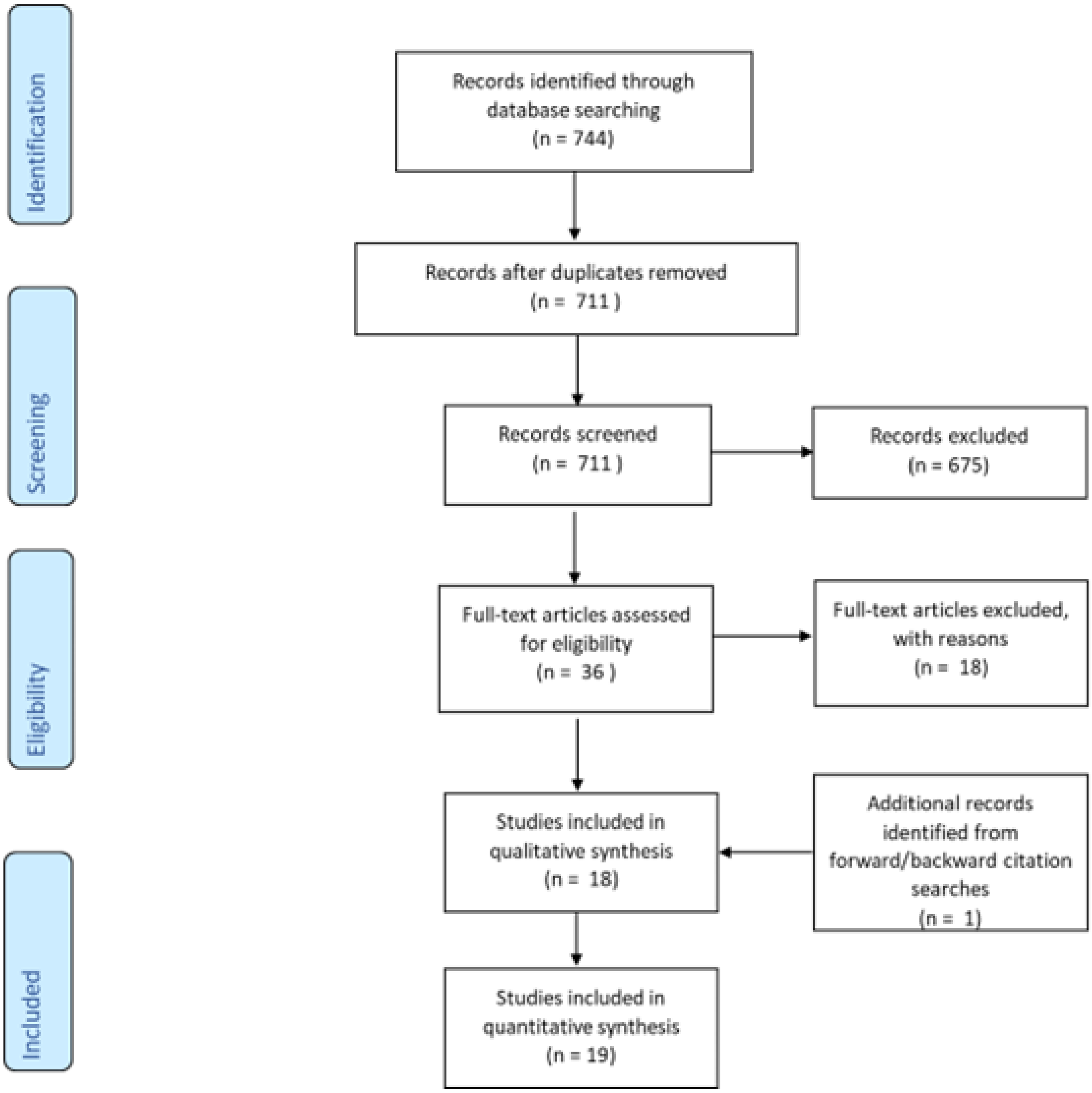
Identification
Screening
Eligibility
Included
Records identified through database searching (n = 744)
Records after duplicates removed (n = 711)
Records screened (n = 711)
Records excluded (n = 675)
Full-text articles assessed for eligibility (n = 36)
Full-text articles excluded, with reasons (n = 18)
Studies included in qualitative synthesis (n = 18)
Additional records identified from forward/backward citation searches (n = 1)
Studies included in quantitative synthesis (n = 19)

Figure 1

PRISMA flow diagram with results for Systematic Review study selection [44].

Note: Reasons for excluding full text articles are included in supplementary data, Table S1

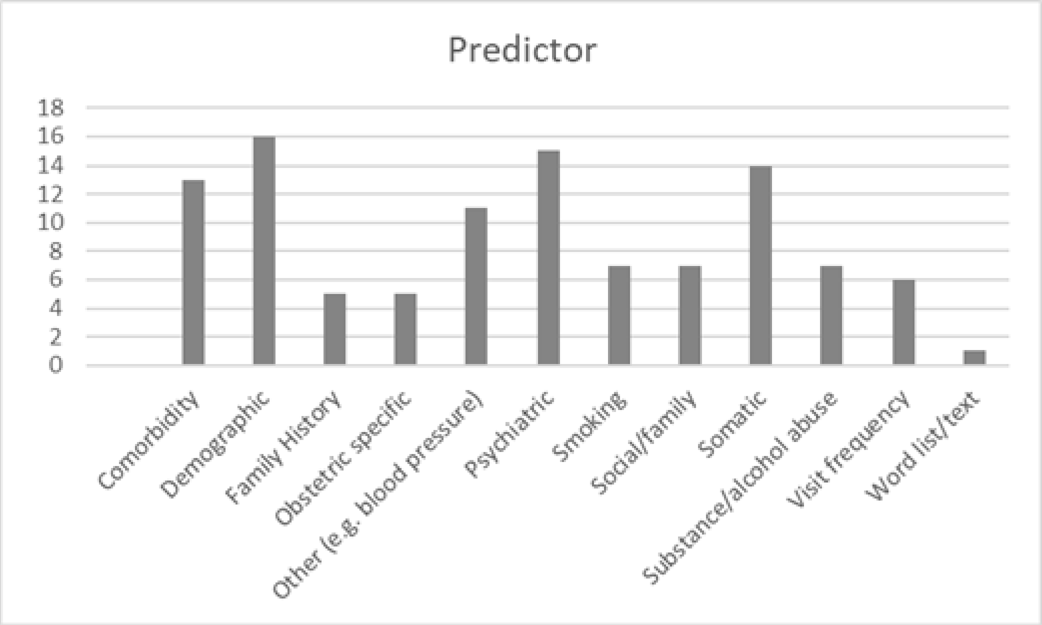

Figure 2

The approximate number of studies using different categories of predictors.

Note 1: Some papers used multiple categories of predictors and not all categorised them.

Note 2: The total number of predictors used was difficult to determine at a summary level as multiple models used different combinations, in some cases no exact number was provided but a reference to a set of definitions used as a starting point.

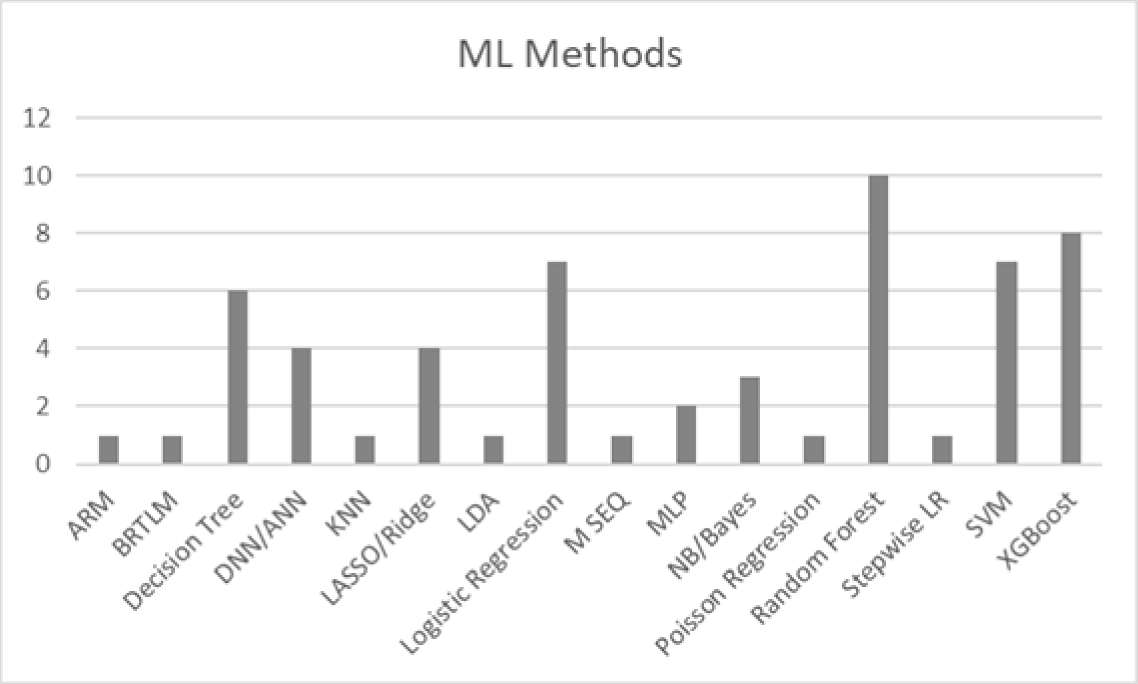

Figure 3

ML/AI Methods for pre-processing and modelling (note LR variants add up to 11). Abbreviations: ANN, Artificial Neural Network; ARM, Association Rule Mining; BRTLM, Bidirectional Representation Learning model with a Transformer architecture on Multimodal EHR; DNN, Deep Neural Network; KNN, K Nearest Neighbours; LASSO, Least Absolute Shrinkage Selection Operator; LR, Logistic Regression; MLP, Multilayer Perceptron; M SEQ, multiple-input multiple-output Sequence; NB, Naïve Bayes; SVM, Support Vector Machine; XGBoost, eXtreme Gradient Boosting.

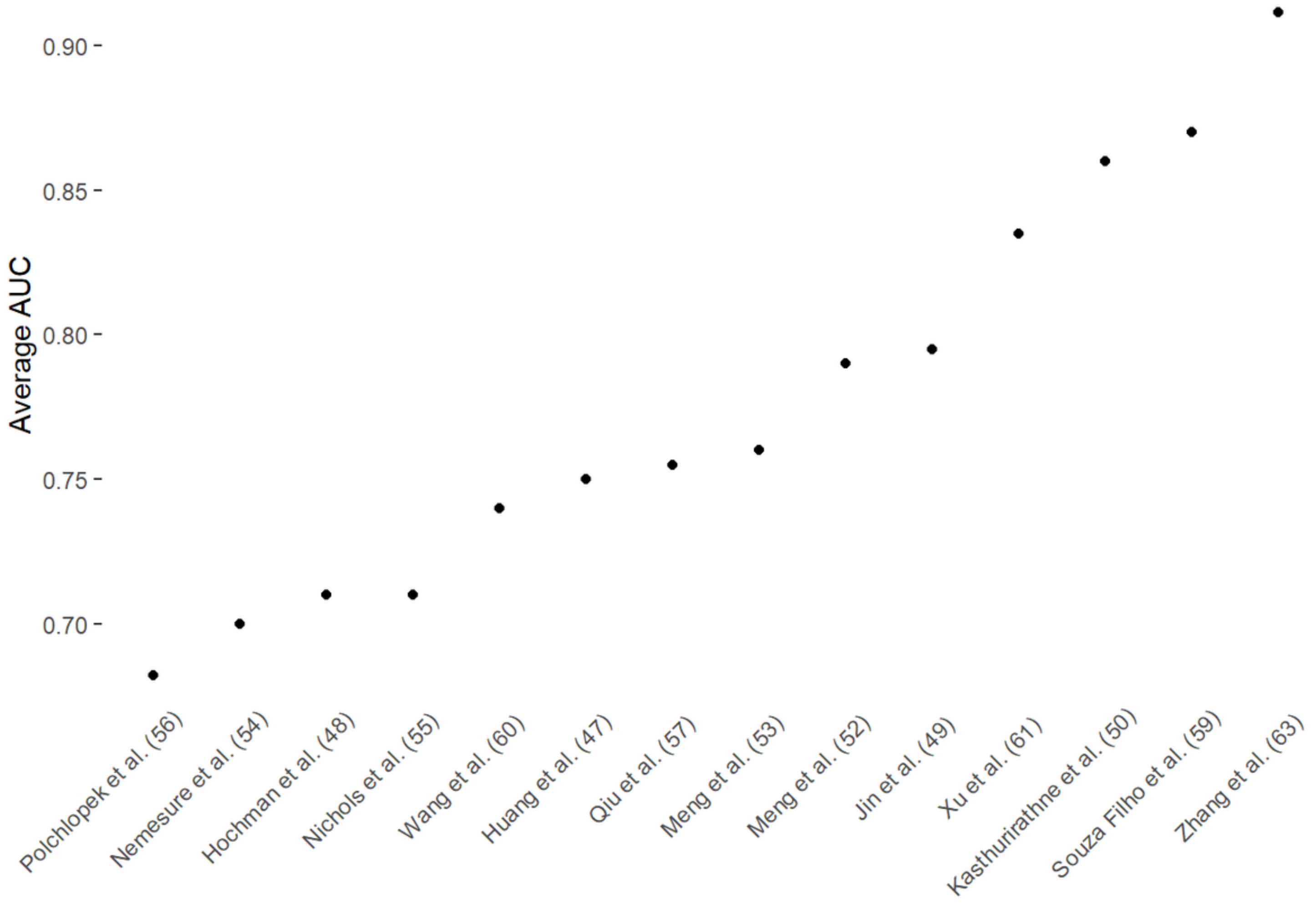

Figure 4

Average AUC performance across studies reporting them (AUC average = 0.78, Standard Deviation AUC Average = 0.07)

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SystematicReviewEHRMLBMCSupplementaryMaterial24012023.docx

- Table4SystematicReviewEHRMLsubmissionBMC25012023.xlsx