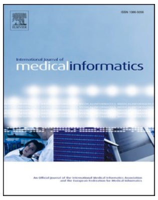
International Journal of Medical Informatics 175 (2023) 105088

Contents lists available at ScienceDirect

International Journal of Medical Informatics

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



Review article

Prediction of disease comorbidity using explainable artificial intelligence and machine learning techniques: A systematic review

Mohanad M. Alsaleh [a,b,*], Freya Allery [a], Jung Won Choi [a], Tuankasfee Hama [a], Andrew McQuillin [c], Honghan Wu [a], Johan H. Thygesen [a]

[a] *Institute of Health Informatics, University College London, London, UK*
[b] *Department of Health Informatics, College of Public Health and Health Informatics, Qassim University, Al Bukayriyah, Saudi Arabia*
[c] *Division of Psychiatry, University College London, London, UK*

ARTICLE INFO

Keywords:
Comorbidity
Multimorbidity
Prediction
Machine learning
Explainable artificial intelligence
Predictive modelling
Systematic review

ABSTRACT

Objective: Disease comorbidity is a major challenge in healthcare affecting the patient's quality of life and costs. AI-based prediction of comorbidities can overcome this issue by improving precision medicine and providing holistic care. The objective of this systematic literature review was to identify and summarise existing machine learning (ML) methods for comorbidity prediction and evaluate the interpretability and explainability of the models.
Materials and methods: The Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) framework was used to identify articles in three databases: Ovid Medline, Web of Science and PubMed. The literature search covered a broad range of terms for the prediction of disease comorbidity and ML, including traditional predictive modelling.
Results: Of 829 unique articles, 58 full-text papers were assessed for eligibility. A final set of 22 articles with 61 ML models was included in this review. Of the identified ML models, 33 models achieved relatively high accuracy (80–95%) and AUC (0.80–0.89). Overall, 72% of studies had high or unclear concerns regarding the risk of bias.
Discussion: This systematic review is the first to examine the use of ML and explainable artificial intelligence (XAI) methods for comorbidity prediction. The chosen studies focused on a limited scope of comorbidities ranging from 1 to 34 (mean = 6), and no novel comorbidities were found due to limited phenotypic and genetic data. The lack of standard evaluation for XAI hinders fair comparisons.
Conclusion: A broad range of ML methods has been used to predict the comorbidities of various disorders. With further development of explainable ML capacity in the field of comorbidity prediction, there is a significant possibility of identifying unmet health needs by highlighting comorbidities in patient groups that were not previously recognised to be at risk for particular comorbidities.

1. Introduction

Disease comorbidity occurs when an individual experiences two or more illnesses simultaneously, which can include physical and/or mental medical conditions [1]. The prevalence of comorbidity is expected to increase significantly in the coming years, with 17% of the UK population projected to have four or more chronic conditions by 2035, nearly double the prevalence of 9.8% in 2015. Moreover, around 67% of those with multiple chronic conditions are expected to have mental illnesses such as dementia, depression, or cognitive impairment [2].

The cost of treating people with multiple long-term conditions is substantial, and it is estimated to rise to over £47 billion by 2035 in the UK alone [3]. Additionally, comorbid patients have a higher mortality rate and poorer quality of life compared to those without comorbidities, making it one of the most pressing issues in healthcare worldwide [4,5]. Also, due to the increased life expectancy, healthcare systems worldwide are facing the challenge of managing the growing incidence of comorbidities [1,6–10].

One potential solution to this challenge is machine learning (ML), a subfield of artificial intelligence (AI). ML models can be trained on large datasets, such as electronic health records (EHRs), to predict the likelihood of a patient developing a particular comorbid condition based on

[*] Corresponding author at: Institute of Health Informatics, University College London, London, UK (Mohanad M. Alsaleh).
E-mail address: mohanad.alsaleh.21@ucl.ac.uk (M.M. Alsaleh).

https://doi.org/10.1016/j.ijmedinf.2023.105088
Received 6 February 2023; Received in revised form 23 March 2023; Accepted 1 May 2023
Available online 4 May 2023

their past medical history, demographics, and other relevant features [11]. ML can also aid in predicting comorbidities by analysing patterns in large datasets of patient data to identify common factors associated with the development of certain comorbid conditions. This can help identify potential risk factors for these conditions, which can then be targeted for prevention or treatment.

However, for ML to be widely adopted in healthcare, it is important to make the models explainable. The use of explainable AI (XAI) methods can help produce an interpretable ML model for disease comorbidity prediction, increasing the transparency of ML [12]. Explainable ML models will allow clinicians to not only forecast future diseases but also understand why an increased risk is predicted. This transparency will be essential for the adoption of ML in healthcare.

There is great potential to improve precision medicine and provide holistic-based care by leveraging ML methods for predicting disease comorbidities. Early and accurate prediction of potential comorbidities can facilitate more efficient treatments and better preventive strategies, resulting in significant cost savings and better health outcomes [13,14]. According to a report by NHS England, providing the same treatment for individuals with the same diseases (one-treatment-for-all) may only be 30–60% effective, and even less effective for people with genetic diseases [15]. By leveraging clinical and genetic data and applying ML with explainable AI methods, healthcare providers can offer more personalised and effective care to patients with comorbidities, ultimately improving treatment and health outcomes [16–18].

2. Significance of study

To the best of our knowledge, there are currently no systematic reviews exploring the literature on the prediction of disease comorbidity using ML methods. Thus, conducting a systematic review on this topic is important as it provides a comprehensive overview of the existing literature in this field. Also, this systematic review allows us to identify gaps in current research, compare the performance of different ML models, and provide a basis for future research. Furthermore, it can facilitate the development of more robust models for disease comorbidity prediction, leading to improved patient outcomes.

This systematic review aimed to identify and summarise existing ML and predictive methods used to predict comorbidities. This study also examined the interpretability and explainability, including XAI methods, of the predictive models where available.

3. Materials and methods

3.1. Case definition for disease comorbidity and predictive modelling

In this systematic review, we have included all papers that contained any of the terms pertaining to comorbidities, such as multimorbidity, comorbid conditions, and multiple conditions, with no restrictions on the type of disease (e.g. rare, genetic and chronic), while maintaining our inclusion and exclusion criteria. This review focused on the various types of ML algorithms and general statistical techniques used in the literature for disease comorbidity prediction. We defined all predictive modelling approaches as ML in this study.

3.2. Search strategy

A systematic literature review was conducted and included all relevant papers prior to March 04, 2023, to examine the use of ML and XAI methods for comorbidity prediction. Three databases (Ovid Medline, Web of Science, and PubMed) were searched using keywords related to explainable AI and ML techniques and comorbidity predictions. We validated our search query by manually identifying relevant publications from PubMed and retrieving them using the search query. To reduce bias and ensure data quality, two reviewers independently screened titles and abstracts and full texts for final inclusion and data extraction. Discrepancies were resolved by consulting a third reviewer and reaching a consensus agreement. The review protocol was registered in PROSPERO (registration number CRD42022332597) to promote transparency and prevent duplication.

The literature search included the following search terms: ("machine learning" OR "artificial intelligence" OR "explainable artificial intelligence" OR "XAI" OR "explainable machine learning" OR "deep learning" OR "data mining" OR "neural network*" OR "association rule mining" OR "pattern analysis" OR "pattern recognition" OR "ensemble learning" OR "statistical learning" OR "support vector machine*" OR "logical learning" OR "Naïve Bayes" OR "Bayesian network*" OR "Gaussian process*") AND ("predict*" OR "prognosis" OR "prognostic" OR "prediction model*" OR "predictive model*") AND ("comorbid condition*" OR "comorbidity*" OR "multimorbidity" OR "multi-morbidity" OR "multimorbid condition*" OR "multi-morbid condition*").

3.3. Inclusion and exclusion criteria

The Population, Intervention, Comparison and Outcome (PICO) framework was used to guide paper selection and develop research questions. The population studied was patients of all ages with comorbidities, the intervention was ML methods for predicting comorbidity, the comparison was different ML algorithms, and the outcome was the type and performance of the ML model and challenges in current research.

"During the title/abstract screening stage, we conducted an independent screening of all identified articles to assess the suitability of the aim and methods of each paper. Specifically, we assessed the aim of the paper by determining whether it addressed the use of machine learning/artificial intelligence in predicting or modelling comorbidity/multimorbidity. We also assessed the methods of the paper by evaluating the use of machine learning/artificial intelligence techniques and the inclusion of relevant statistical analyses. Additionally, we considered the presence of our main keywords: machine learning/artificial intelligence, comorbidity/multimorbidity, and prediction/predictive modelling.

During the title/abstract screening stage, an independent screening of all identified articles was conducted based on their eligibility for inclusion in this review. This was determined by assessing whether the aim and methods in the abstracts fell within the scope of this review and by considering the presence of the main keywords: machine learning/artificial intelligence, comorbidity/multimorbidity, and prediction/predictive modelling in the title and abstract.

Specifically, English-language, peer-reviewed papers recruiting patients with a certain disease to predict comorbid condition(s) of the disease being studied, with no publication year restrictions were included. Studies without sample size or key sample characteristics, those that predicted mortality, readmission or drug side effects without predicting comorbidity, and those without information on ML model performance metrics were excluded, along with review and overview studies.

3.4. Extraction and analysis

The Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) framework was used to report the findings of the systematic review (Fig. 1 and Supplementary Table 1) [19]. To increase comprehensiveness, the literature search included manual searches of references and citations in the selected studies and related articles in Google Scholar. Two reviewers extracted information from the selected articles, considering study and sample characteristics, source of data, primary disease being studied, comorbidities predicted, ML algorithms used, model interpretation and explainability, and key findings limitations. The findings were synthesised using a narrative approach with plots, figures, and tables summarising the results.

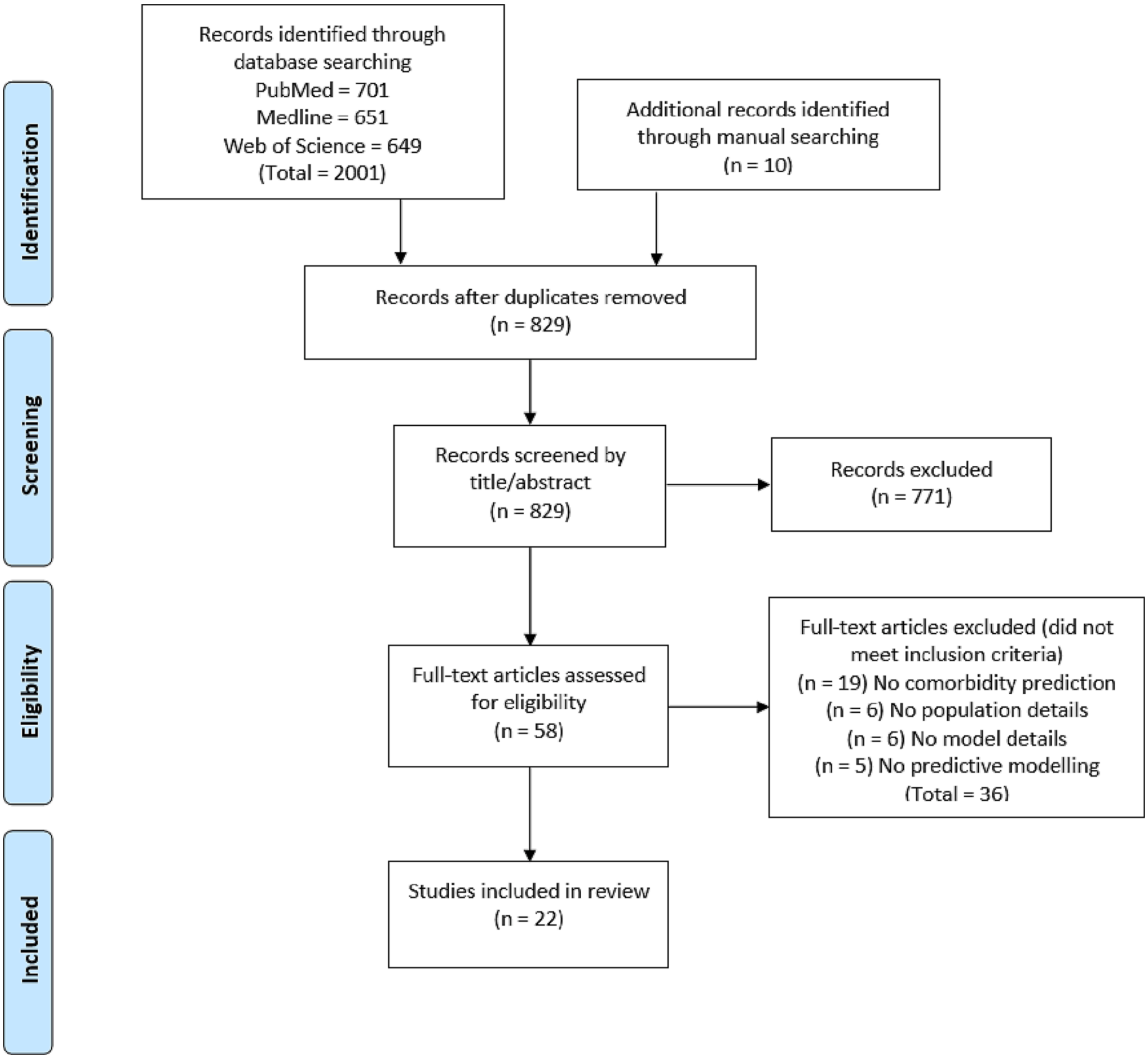

Identification
Screening
Eligibility
Included
Records identified through database searching
PubMed = 701
Medline = 651
Web of Science = 649
(Total = 2001)
Additional records identified through manual searching
(n = 10)
Records after duplicates removed
(n = 829)
Records screened by title/abstract
(n = 829)
Records excluded
(n = 771)
Full-text articles assessed for eligibility
(n = 58)
Full-text articles excluded (did not meet inclusion criteria)
(n = 19) No comorbidity prediction
(n = 6) No population details
(n = 6) No model details
(n = 5) No predictive modelling
(Total = 36)
Studies included in review
(n = 22)

Fig. 1. PRISMA flow diagram for the systematic review. PRISMA Preferred Reporting Items for Systematic Reviews and Meta-Analysis.

3.5. Risk of bias and quality assessment

The Prediction model study Risk of Bias Assessment Tool (PROBAST) was used to assess the Risk of Bias (ROB) and applicability of the best-performing ML model in the selected articles [20]. PROBAST contains 20 questions divided into 4 domains (participants, predictors, outcome, analysis), with each domain scored as low, unclear or high risk of bias. It also assesses the study's applicability to the target population, predictors, and outcomes. Assessment of the PROBAST questions was done by one reviewer for all the selected articles. The authors' main model was considered for assessment and comparison, or if not specified, the model with the highest accuracy and/or highest AUC value was chosen. This approach facilitated a more direct ROB assessment and comparison of the ML models across the identified studies, particularly between studies with multiple trained models and those with only one model. The ROB assessment considered: 1) Participants' data sources and selection, 2) Predictors' definition and availability, 3) Outcome definition and 4) Analysis: participants, predictors, missing data and model performance.

4. Results

4.1. Main findings

A total of 2001 articles were found across searches in Ovid Medline, Web of Science and PubMed, including articles found through manual searches by looking into the references of identified key papers. After removing duplicates, 829 studies were assessed for inclusion. Of 829 articles, 58 full-text articles were assessed to determine eligibility. A total of 22 articles met the inclusion criteria and were included in this review (Fig. 1). The most frequent exclusion criteria were: (1) comorbidity data was used to predict mortality and/or disease severity, rather than comorbidity; and (2) the studies lacked information on model development or population characteristics. The identified studies were published between 2009 and 2023 (median = 2020) focusing on a variety of diseases to predict their associated comorbidities using ML algorithms. Most of the studies were conducted in the USA (n = 8), followed by Australia (n = 6). Across the 22 studies, a wide variety of ML algorithms (n = 61) were used for disease comorbidity predictions (see Supplementary Table 2 for more details on each study).

4.2. Studies and disease comorbidity

Most of the studies (n = 14) focused on predicting comorbidities of chronic conditions, such as diabetes, heart disease, and hypertension. A few studies (n = 5) aimed to predict comorbidities of neurological and mental health conditions, such as epilepsy, bipolar disorder and depression. Only one study investigated the comorbidities of a genetically rare condition (tuberous sclerosis complex) [21]. The authors incorporated genetic data by identifying which patients had genetic mutations in genes TSC1 and TSC2 [21]. The comorbidities that these studies predicted were diverse due to the primary “recruitment” diseases and characteristics of the participants [14,22–27].

Across the 22 studies, over 30 unique comorbidities were analysed and predicted for multiple medical conditions ranging from 1 to 34 with an average value of 6. We observed that diabetes and depression were among the top targeted comorbidities to be predicted (Table 1). Identifying such comorbid conditions facilitated the authors to build a disease network that could enhance the understanding of the diseases’ mechanisms and their relationships [25,18,28,29,30]. Several studies analysed the comorbidities in individuals with multiple chronic conditions, focusing on the connections between diseases and the ability of ML to predict the comorbidities [23,29,30].

4.3. Data sample size and types

The sample sizes for the 22 studies varied (mean = 52,724, median = 6,883), with the lowest sample size being 77 participants and the highest sample size being 257,633 [21,23]. The majority of the studies (n = 19)—all retrospective cohort and/or case-control studies—used data from databases, such as EHRs, registries and corporate databases, while three studies used data obtained from interviews and questionnaires [27,31,32]. Of the 19 studies, 7 studies used administrative and claim-based datasets (five of them shared the same data source; the Commonwealth Bank Health Society (CBHS) in Australia) to forecast comorbid conditions using ICD-10-based diagnoses [25,29,30,14,28].

4.4. ML methods and features

The most used ML algorithms for predicting comorbidities in the reviewed studies were logistic regression (n = 11) and random forest (n = 10), followed by support vector machines (n = 5) and Decision Trees (n = 4). Various deep learning algorithms such as neural networks, wide & deep learning, multi-layer perceptron and convolutional neural networks were used in 5 studies [14,26–30,33], while 2 probabilistic ML algorithms (Bayesian networks and Naive Bayes) were applied [29,34,25,23]. Three different regression techniques were observed, Poisson regression, rigid logistic regression and multinomial logistic regression [35,36,37]. Some authors took an unsupervised learning approach to identify clusters and relationships of diseases, which were then used to build a disease comorbidity network [25,30,23,38].

The selected studies analysed various features from various sources. The most common features across all studies were: (a) demographic information such as age, gender, race, and marital status; (b) clinical information such as diagnosis codes, genes involved, and family history and (c) health registry information such as hospital diagnoses, drug prescriptions, body mass index (BMI) and smoking status. These features were used to create predictive models to predict comorbidities. The number of features varied between studies, ranging from 6 to 22. Additionally, the study also considered network features, such as centrality measures and clustering coefficient, in some studies [29,30,25]. The Python programming language was used in 46% of the studies, followed by R (41%) and MATLAB (13%). Each study employed different modelling techniques and predictors.

4.5. Risk of bias

An overall assessment of ROB for each study’s best-performing model was performed based on the four domains of PROBAST: participants, predictors, analysis, and outcomes domains. The participants domain had the highest number of studies showing a high ROB (n = 8) due to the use of questionnaires/interviews and bias in participants selection (e.g. study design and the inclusion/exclusion criteria), which might result in a study population that is not unrepresentative of the target population [27,31,32]. The Analysis domain had unclear ROB (n = 8) due to a lack of information on missing data handling and predictor selection. The predictors domain was second most highly rated as having a high ROB (n = 7), due to their unavailability at the time the model was used, while the outcome was defined and measured differently for the cases and controls, indicating either a high and/or unclear ROB (n = 8). Overall, 72% of the studies exhibited high and/or unclear concerns regarding their risk of bias.

With regards to the applicability, 45% of the studies showed a high and/or unclear concern. While our systematic review aimed to evaluate relevant studies on ML predictions of comorbidity regardless of the type of data, it was observed that the included studies utilised a variety of data sources and recruitment methods, which may affect potential bias and the general applicability of these studies. Studies using interviews and questionnaires to recruit participants may have limited applicability to the broader population due to potential biases and limitations related to self-reporting or recruitment of specific subsets [27,31,32]. Overall, a total of 55% of the studies demonstrated a low concern regarding their applicability (Fig. 2 & Supplementary Table 4).

Table 1
Most predicted comorbidities found in the studies.

Study	Comorbidity				
	Diabetes	Depression	Heart disease	Hypertension	Anxiety
Faruqui et al. (2018)		✓			
Wang et al. (2021)		✓			
Lu and Uddin (2022)	✓	✓	✓	✓	
Jin et al. (2015) *		✓			
Tennenhouse et al. (2020)		✓			✓
Glauser et al. (2020)		✓			✓
Jin et al. (2015) *		✓			
Linden et al. (2021)	✓				✓
Farran et al. (2013)	✓			✓	
Ojeme & Mboghoet (2016)	✓		✓	✓	
Abdalrada et al. (2022)	✓		✓		
Uddin et al. (2022)	✓		✓		
Dworzynski et al. (2020)			✓		
Nikolaou et al. (2021)				✓	
Lu and Uddin (2023)	✓	✓	✓	✓	
Chari et al. (2023)	✓				
Khan et al. (2019)	✓				
Frequency	9	8	6	5	3

All comorbidities can be found in Supplementary Table 2.

* Different studies. Refer to the references for the title and full author details.

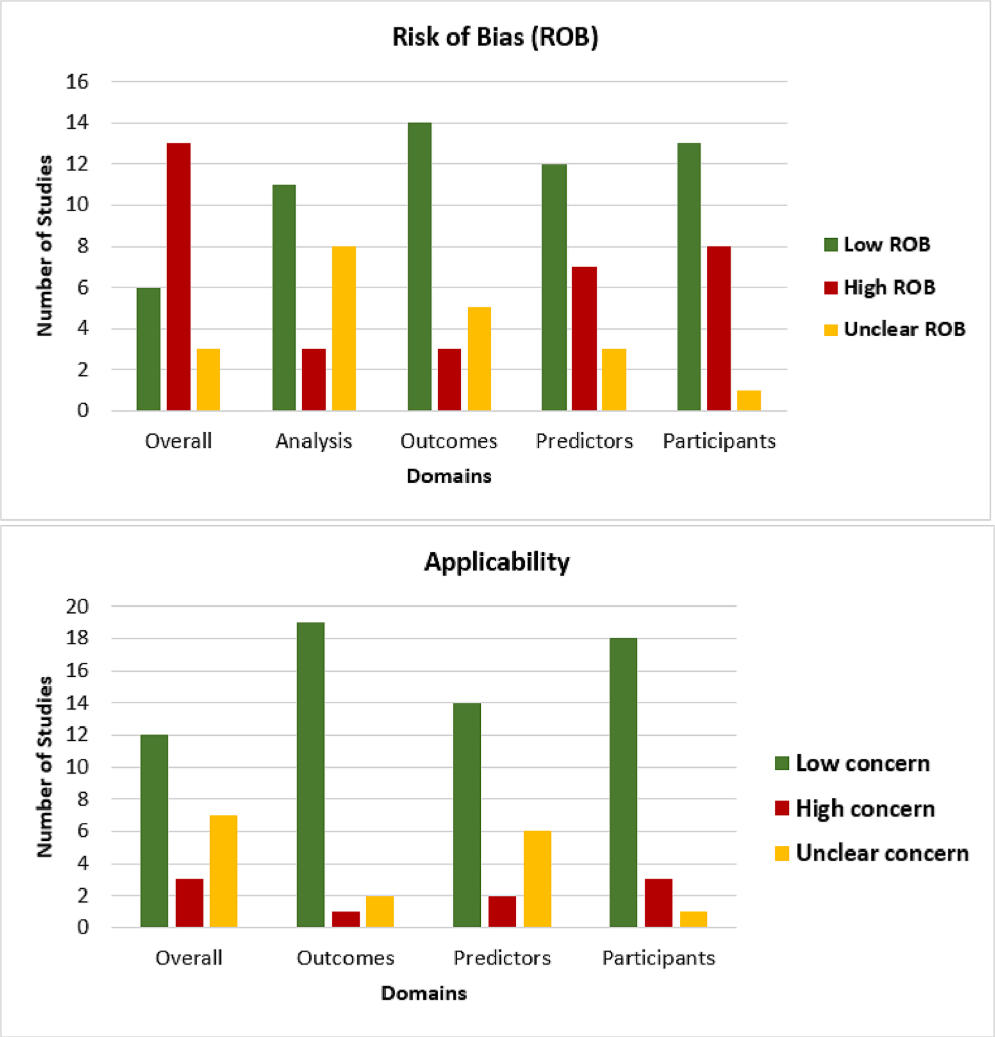


Fig. 2. Study's best-performing model ROB and applicability assessed by PROBAST. Green Indicates a low ROB/low concern for applicability, red a high ROB/high concern for applicability and yellow an unclear ROB/unclear concern for applicability. ROB = risk of bias; PROBAST = Prediction model study Risk of Bias Assessment Tool. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

4.6. Model performance and validation

The majority of the studies (n = 15) used AUC to evaluate the performance of the developed ML models, while 7 studies did not report AUC but instead reported accuracy and the Hamming score [31,29,35,39,28,40,41]. The best-performing models in the studies achieved an acceptable to high-performance score (80–95%). Neural networks (NN) and Poisson regression (PR) algorithms had the highest AUC score, followed by the decision tree algorithm, compared to the others (0.89, 0.89 and 0.88, respectively) (Fig. 3) [27,36,38].

The XGBoost algorithm achieved the highest accuracy score (95.05%), followed by CART (94.09%) and convolutional neural network (91.7%) in the studies reporting accuracy [29,35,39]. Across all studies, only one study used the Hamming score to evaluate the performance of the model (Hamming score = 0.91) (see Supplementary Table 3 for all ML models' performances). Most studies used k-fold cross-validation (n = 16) as a validation technique, while others used external validation (n = 3) [14,35,34], a validation set of the original dataset (n = 4) [21,30,33,22] and baseline models in the literature (n = 2) [23,26]. Some used multiple validation methods.

4.7. Evaluation of interpretability and explainability of models

Out of the 22 studies analysed, only five utilised explainability techniques to render their ML models interpretable. These methods included neural networks, XGBoost (used twice), multi-layer perceptron and random forest, which achieved AUC scores ranging from 0.73 to 0.83 and accuracy levels between 90% and 95% [14,29,28,22]. Two studies employed an explainability approach for XGBoost to predict comorbid conditions of chronic diseases such as heart disease and diabetes, using the built-in feature importance attribute in the algorithm that estimates feature importance based on weight, gain, and coverage [28,29]. Uddin et al. (2022) identified the number of episodes, alcohol and network features such as degree centrality, transitivity and PageRank score as the most important features, with F1 feature importance gain scores higher than 10 [29]. Meanwhile, Lu and Uddin (2023) highlighted network features such as degree centrality and Jaccard coefficient as significantly impacting XGBoost performance, with feature importance greater than 0.1 [28]. Zhang-James et al. (2020) followed a similar approach but used a different model (i.e. random forest) to identify the most critical features for predicting substance use disorder among children with attention-deficit/hyperactivity disorder. The top important features in the RF model included teenage criminal records (from onset age 15 up to age 17), childhood (age 2–12) ADHD diagnosis,

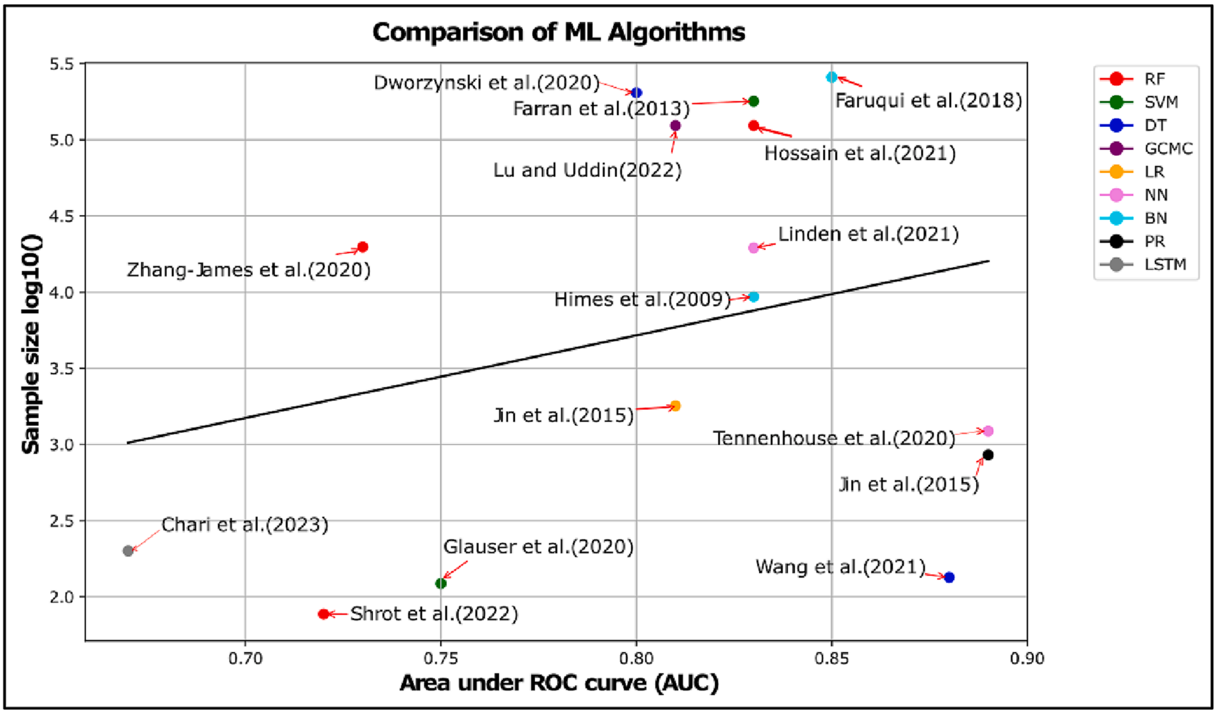


Fig. 3. Comparison of the best-performing ML algorithms found in the studies by AUC and sample size. This scatterplot shows each data point represented by a dot, with the dot colour corresponding to the algorithm used, and the study name and year as a label close to the dot. A trendline (in a dashed black line) fitted to the data using linear regression. The y-axis is transformed using log10. The legend shows the ML algorithm names with their corresponding dot colours. SVM support vector machines, NN neural network, CNN convolutional neural network, DT decision tree, RF random forest, GCMC graph convolutional matrix completion, LR logistic regression, BN Bayesian network, PR Poisson regression, LSTM long-short term memory.

stimulant treatment before the age of 18 and diagnosis of anxiety disorder [22].

Two other studies used neural network and multi-layer perceptron models, respectively, to predict the comorbidities of epilepsy and diabetes, employing an XAI method called Shapley Additive Explanations (SHAP)—based on the contribution and relevance of each feature to the prediction—to make their models explainable [14,33]. The most relevant features in the neural network model included age, gender, drug prescriptions (e.g., topiramate) and hypertension diagnosis [14], while the multi-layer perceptron model identified age at onset and conditions related to circulatory, musculoskeletal and respiratory systems as the most important features [33].

5. Discussion

To the best of our knowledge, this systematic review is the first to examine the use of ML and XAI for predicting comorbidities. Such technologies can help mitigate the risk and burden of comorbidities by predicting them, thus improving quality of life and care [39,30,42]. Most comorbidities predicted were already listed in existing comorbidity indices such as Elixhauser Comorbidity Index [43]. Studies have shown that diabetes, hypertension and depression are among the ten most prevalent comorbidities associated with morbidity [44,45,46]. The results of this review have also shown that these conditions were the most commonly analysed and predicted comorbidities across the identified papers.

The ability to predict comorbidities may vary based on the primary disease being studied [23,25]. Some diseases may have more complex relationships with comorbidities than others, and patients with these conditions could have different demographic and clinical profiles [8,4]. Additionally, some comorbidities may be more strongly associated with the primary disease, while others may be more variable across patient populations. This will impact the ability and performance of ML algorithms.

The different methods used to capture phenotypes play a significant role in the ability to predict comorbidities. For example, ICD10 codes are commonly used in clinical settings but may not capture all relevant information, while phenotyping algorithms (e.g. the CALIBER and HDR-UK Phenotype Library) and clinical trial definitions may be more specific but require additional validation [49,50]. The potential biases associated with different methods of phenotype capture may be difficult to evaluate and compare between studies with variable approaches. For example, some methods may be more likely to capture certain types of patients or exclude others, which could affect the ability of the ML models [51,52]. Thus it may be necessary to develop specialised models or to use additional clinical and genetic information to improve predictions. Further delineation of the impact of phenotyping methods on prediction could be an interesting subject for further studies.

It was observed that the association between sample size and model performance appeared to be positive, indicating that increasing the sample size improves the performance of the model. In general, having a larger sample size can lead to better model performance because it provides more data for the model to learn from, which can help to increase the model's performance and lead to better generalisation and a decrease in overfitting [47,48]. However, it is important to note that this relationship is not always straightforward, as other factors such as the quality of the data, the complexity of the model, and the presence of outliers can also affect model performance.

The ML algorithms identified in this review showed the capability of processing vast amounts of data from various sources, such as EHRs, to identify patterns and associations by building comorbidity networks that may not be evident to human experts [29,25,30]. This can lead to more accurate and personalised predictions of comorbidities, which can help clinicians in making informed decisions about patient care. However, there are also some limitations to using ML algorithms in predicting comorbid conditions. ML algorithms are only as good as the data they are trained on, so if the data used to train the algorithm is biased or incomplete, the algorithm may make inaccurate predictions [35,27]. Additionally, ML algorithms can be complex and difficult to interpret, which can make it challenging for clinicians to understand how the algorithm arrived at its predictions [31]. Therefore, it is essential to ensure that ML algorithms are validated and transparently communicated to healthcare providers to avoid the potential for incorrect or misleading predictions. To assist future studies in identifying the best-performing models published to date, we have highlighted the best-performing model in each study with bold text and an asterisk in the Supplementary Table 2.

The AUC and accuracy are common metrics for evaluating a predictive model. Most studies used AUC (n = 15), as AUC calculates True Positive Rate (TPR), False Positive Rate (FPR), sensitivity, and specificity, making it less biased than accuracy, which only calculates correct predictions. Neither the majority class nor the minority class is subject to bias when using AUC. Thus, AUC can be a better indicator of model performance, especially when dealing with imbalanced data, as it avoids bias towards both the majority and minority classes [53–55]. To address the class imbalance, one study used up- and down-sampling to increase the accuracy of random forest [35].

The authors in the selected papers focused on applying ML to predict comorbidities, using pre-existing lists of comorbidities, with a limited scope of comorbidities studied (mean = 6). These findings suggest that no studies are coming close to a high level of inclusion. The lack of genetic data and phenotyping methods hindered their ability to uncover novel findings. The complexity of identifying comorbidity vs. primary disease and the overlap of symptoms and signs among diseases were also challenges faced by the authors [56,57].

The research on comorbidity analysis has seen an increase in activities in recent years, with most of the chosen studies (n = 15) being carried out between 2020 and 2023. This indicates a growing interest in the application of ML in predicting comorbidities for improving patient care and treatment outcomes. The studies also highlight the need to move away from a single-disease approach towards a more holistic approach to patient care [58]. While most studies focused on common diseases and their associated comorbidities, one study explored comorbidities in a rare medical condition, tuberous sclerosis complex, using both EHR and genetic data [21]. Incorporating genetic data in the analysis of comorbidities can reveal new insights into disease mechanisms by identifying genotype-phenotype associations. This approach can help in understanding the comorbidities in carriers of pathogenic mutations and is strongly encouraged [59,60].

In this review, the interpretability and explainability of ML models were explored. The interpretability was mainly achieved by estimating the feature importance, but there was no standard evaluation method for explainable ML models, hindering a fair comparison. These findings highlight the need for a standard evaluation method and the limitations of current XAI methods in aiding understanding and comparability across domains. Although there is a trade-off between model explainability and performance, the explainable ML identified in this review achieved an acceptable to high performance [14,29,28]. Since XAI is regarded as an emerging field in AI, increased interdisciplinary collaboration between various professionals, including clinicians and engineers, is necessary so that more robust, trustworthy and explainable AI models are built. However, the effectiveness of explainability mostly depends on how human users understand the ML model [61]. Although explainable models may achieve comparable performance to non-explainable models, it is vital to consider factors such as the number and type of variables, and the target class accuracy to accurately evaluate the performance of explainable models, as these factors may impact the performance of explainable ML, as well as non-explainable ML models making direct comparisons difficult [62].

While our analysis of the included studies demonstrated promising results in the use of machine learning models for predicting comorbidity, it is important to note that certain confounding variables, such as age or socio-economic status, may not have been fully controlled in the selected studies. Additionally, a lack of diversity in the study populations was noted, with the majority of participants being of a similar race/ethnicity or geographic location. This lack of diversity, combined with a lack of external validation, could impact the generalizability and performance of the machine learning models.

There is a gap in the literature on how ML and explainability techniques could be used to simultaneously identify comorbidities of diseases using both genetic and EHR data. Regarding methodological limitations and gaps, a lack of genetic data such as genetic mutations and a comprehensive list of comorbidities was identified. This could be due to a lack of access to databases containing both genetic and EHR data. Thus, there is a need for a more comprehensive list of comorbidities and more explainable ML models that exploit both genetic and EHR data.

Based on the knowledge obtained from this review, we believe that clinicians and researchers developing ML models for the prediction of comorbidity in medical care can take several steps to enhance their impact. These include; (a) incorporating clinical expertise and stakeholder input throughout all stages of model development and validation, (b) standardising approaches to data collection, curation, phenotyping and model validation to allow for replication and informed comparison between studies, (c) building transparent and explainable predictive models to enhance clinical usability and evaluation of bias, (d) integrating other types of data such as genetic to enhance performance, and lastly (e) promoting open data sharing and collaboration and considering ethical and legal implications of newly developed models [63,64]. These steps can ensure that ML predictions of comorbidity are more clinically relevant and address the needs of patients and healthcare providers, while also maintaining patient privacy and avoiding exacerbating health disparities [65]. We believe that these suggestions could help to advance the field of ML predictions of comorbidity and ultimately improve patient outcomes.

5.1. Limitations and future work

This study had several limitations. First, although some studies reported precision, recall, sensitivity and F1, only AUC and accuracy scores were considered for reporting model performance. Although AUC and accuracy are commonly used metrics for evaluating machine learning models, their suitability depends on the specific predictive task at hand. Other metrics such as precision, recall, sensitivity, specificity, and F1-score may be more appropriate in some cases. Therefore, researchers should choose evaluation metrics based on the specific context and goals of their study to ensure the most accurate assessment of model performance.

Second, even though the AUC provides a valid approach for the ML model comparison, not all studies reported AUC values. It is therefore difficult to fairly compare the ML models using two different performance evaluation metrics (i.e. AUC and accuracy). Also, we did not specifically investigate whether the included studies used other metrics to ensure appropriate prediction of comorbidity prevalence. While we acknowledge that this is an important issue, it was beyond the scope of our review. Therefore, future research could explore the use of additional metrics to improve the accuracy of comorbidity prediction.

Another limitation of our systematic review is that not all studies reported AUC values, which can make it difficult to compare the performance of machine learning models fairly when using different evaluation metrics, such as AUC and accuracy. In addition, we did not investigate whether the included studies used other metrics to ensure appropriate prediction of comorbidity prevalence, which could be relevant to explore in future research. Since most of the included studies had developed multiple ML models, we only compared the best-performing ML model identified in each study. As a result, we only assessed those ML models in terms of the ROB.

Although our search query was validated by manually identifying and retrieving relevant publications from PubMed, other relevant studies predicting comorbidities in patients may have been missed due to variations in search terms (i.e. relevant papers may not have been captured by the keywords used in the literature search) or ambiguity in the studies' titles, aims and methods. Future research could broaden the search criteria to include additional studies that may have been missed in our review.

Based on this study outcome, we highly recommend studies to analyse the variations in model accuracy among various subgroups in the sample, taking into account their unique characteristics, such as age, gender, ethnicity and socioeconomic status. Future studies should also aim to include more diverse study populations and control for potential confounding variables, to ensure the accuracy and applicability of the developed models in a broader range of settings. Moreover, to further expand the scope of the investigation, we suggest exploring the effects of integrating additional factors, such as social determinants of health and genetic traits, on the accuracy of the predictions in future reviews.

Future studies should prioritise the development and validation of these models on larger and more diverse populations, as this can enhance the generalizability and applicability of ML models for comorbidity prediction. Additionally, it is recommended that future research take a disease-agnostic approach and consider a broader range of conditions beyond the specialty of the disease under examination. With the aid of EHRs and genetic data, along with validated phenotype libraries such as CALIBER and HDR-UK Phenotype Library [49,50], the future of comorbidity analysis holds promise in identifying and analysing hundreds, if not thousands, of diseases.

6. Conclusion

In conclusion, the use of AI methods has shown great potential in improving the quality and cost-effectiveness of healthcare, as demonstrated by its use in predicting comorbidities in individual patients. While a wide range of ML algorithms was identified, only a few were made explainable, highlighting the need for further development in this area. With continued research, there is a significant possibility of boosting precision medicine by identifying unmet health needs in patient populations not previously known to be at risk for specific comorbidities. Furthermore, the integration of additional variables such as social determinants of health and genetic characteristics could lead to even more accurate and personalised predictions. Overall, the transparent and explainable nature of AI has the potential to revolutionise the way we approach disease prevention and treatment, ultimately leading to improved patient outcomes and better healthcare for all.

7. Authors' contributions

MMA and JHT formulated the research topic and questions, MMA conducted the literature search, summarised the findings and drafted the manuscript. MMA, FA and TH independently screened the identified articles. MMA, JHT, AM, HW, FA, JWC and TH reviewed the manuscript and provided comments and clarifications.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ijmedinf.2023.105088.

References

[1] C. Harrison, et al., Comorbidity versus multimorbidity: Why it matters, J. Multimorb. Comorb., 11, 2633556521993993 (2021).

[2] A. Kingston, et al., Projections of multi-morbidity in the older population in England to 2035: estimates from the Population Ageing and Care Simulation (PACSim) model, Age Ageing 47 (2018) 374–380.

[3] J. Appleby, Spending on health and social care over the next 50 years, Why think long term ? Spending on health and social care over the next 50 years. Why think long term ? https://pascal-francis.inist.fr/vibad/index.php?action=getRecordDetail&idt=26923988 (2013).

[4] E. Ge, Y. Li, S. Wu, E. Candido, X. Wei, Association of pre-existing comorbidities with mortality and disease severity among 167,500 individuals with COVID-19 in Canada: A population-based cohort study, PLoS One 16 (2021) e0258154.

[5] Y.K. Lee, et al., The relationship of comorbidities to mortality and cause of death in patients with differentiated thyroid carcinoma, Sci. Rep. 9 (2019) 11435.

[6] J.F. Figueroa, et al., International comparison of health spending and utilization among people with complex multimorbidity, Health Serv. Res. 56 (Suppl 3) (2021) 1317–1334.

[7] S.I. Cho, S. Yoon, H.-J. Lee, Impact of comorbidity burden on mortality in patients with COVID-19 using the Korean health insurance database, Sci. Rep. 11 (2021) 6375.

[8] D. Sarfati, B. Koczwara, C. Jackson, The impact of comorbidity on cancer and its treatment, CA Cancer J. Clin. 66 (2016) 337–350.

[9] J.F. Piccirillo, I. Costas, The impact of comorbidity on outcomes, ORL J. Otorhinolaryngol. Relat. Spec. 66 (2004) 180–185.

[10] R. Gijsen, et al., Causes and consequences of comorbidity: A review, J. Clin. Epidemiol. 54 (2001) 661–674.

[11] J. Jovel, R. Greiner, An Introduction to Machine Learning Approaches for Biomedical Research, Front. Med. 8 (2021), 771607.

[12] A.M. Antoniadi, et al., Current Challenges and Future Opportunities for XAI in Machine Learning-Based Clinical Decision Support Systems: A Systematic Review, NATO Adv. Sci. Inst. Ser. E Appl. Sci. 11 (2021) 5088.

[13] A. Kline, et al. Multimodal Machine Learning in Precision Health. *arXiv [cs.LG]* (2022).

[14] T. Linden, et al., An Explainable Multimodal Neural Network Architecture for Predicting Epilepsy Comorbidities Based on Administrative Claims Data, Front. Artif. Intell. 4 (2021) 610197.

[15] England, N. H. S. Improving outcomes through personalised medicine. *NHS England* https://www.england.nhs.uk/wp-content/uploads/2016/09/improving-outcomes-personalised-medicine.pdf (2016).

[16] J. Zhao, et al., Learning from Longitudinal Data in Electronic Health Record and Genetic Data to Improve Cardiovascular Event Prediction, Sci. Rep. 9 (2019) 717.

[17] J. Deng, T. Hartung, E. Capobianco, J.Y. Chen, F. Emmert-Streib, Editorial: Artificial Intelligence for Precision Medicine, Front. Artif. Intell. 4 (2021) 834645.

[18] P. Akram, L. Liao, Prediction of comorbid diseases using weighted geometric embedding of human interactome, BMC Med. Genomics 12 (2019) 161.

[19] M.J. Page, et al., The PRISMA 2020 statement: An updated guideline for reporting systematic reviews, J. Clin. Epidemiol. 134 (2021) 178–189.

[20] R.F. Wolff, et al., PROBAST: A Tool to Assess the Risk of Bias and Applicability of Prediction Model Studies, Ann. Intern. Med. 170 (2019) 51–58.

[21] S. Shrot, et al., Prediction of tuberous sclerosis-associated neurocognitive disorders and seizures via machine learning of structural magnetic resonance imaging, Neuroradiology 64 (2022) 611–620.

[22] Y. Zhang-James, et al., Machine-Learning prediction of comorbid substance use disorders in ADHD youth using Swedish registry data, J. Child Psychol. Psychiatry 61 (2020) 1370–1379.

[23] S.H.A. Faruqui, A. Alaeddini, C.A. Jaramillo, J.S. Potter, M.J. Pugh, Mining patterns of comorbidity evolution in patients with multiple chronic conditions using unsupervised multi-level temporal Bayesian network, PLoS One 13 (2018) e0199768.

[24] B. Farran, A.M. Channanath, K. Behbehani, T.A. Thanaraj, Predictive models to assess risk of type 2 diabetes, hypertension and comorbidity: machine-learning algorithms and validation using national health data from Kuwait–a cohort study, BMJ Open 3 (2013).

[25] M.E. Hossain, S. Uddin, A. Khan, Network analytics and machine learning for predictive risk modelling of cardiovascular disease in patients with type 2 diabetes, Expert Syst. Appl. 164 (2021) 113918.

[26] H. Jin, S. Wu, P. Di Capua, Development of a Clinical Forecasting Model to Predict Comorbid Depression Among Diabetes Patients and an Application in Depression Screening Policy Making, Prev. Chronic Dis. 12 (2015) E142.

[27] L.G. Tennenhouse, R.A. Marrie, C.N. Bernstein, L.M. Lix, & CIHR Team in Defining the Burden and Managing the Effects of Psychiatric Comorbidity in Chronic Immunoinflammatory Disease. Machine-learning models for depression and anxiety in individuals with immune-mediated inflammatory disease, J. Psychosom. Res. 134 (2020) 110126.

[28] H. Lu, S. Uddin, Embedding-based link predictions to explore latent comorbidity of chronic diseases, Health Inf. Sci. Syst. 11 (2023) 2.

[29] S. Uddin, et al., Comorbidity and multimorbidity prediction of major chronic diseases using machine learning and network analytics, Expert Syst. Appl. 205 (2022) 117761.

[30] H. Lu, S. Uddin, A disease network-based recommender system framework for predictive risk modelling of chronic diseases and their comorbidities, Appl. Intell. 52 (2022) 10330–10340.

[31] B. Ojeme, A. Mbogho, Selecting Learning Algorithms for Simultaneous Identification of Depression and Comorbid Disorders, Procedia Comput. Sci. 96 (2016) 1294–1303.

[32] T. Glauser, et al., Identifying epilepsy psychiatric comorbidities with machine learning, Acta Neurol. Scand. 141 (2020) 388–396.

[33] S. Chari, et al., Informing clinical assessment by contextualizing post-hoc explanations of risk prediction models in type-2 diabetes, Artif. Intell. Med. 137 (2023) 102498.

[34] B.E. Himes, Y. Dai, I.S. Kohane, S.T. Weiss, M.F. Ramoni, Prediction of chronic obstructive pulmonary disease (COPD) in asthma patients using electronic medical records, J. Am. Med. Inform. Assoc. 16 (2009) 371–379.

[35] V. Nikolaou, et al., The cardiovascular phenotype of Chronic Obstructive Pulmonary Disease (COPD): Applying machine learning to the prediction of cardiovascular comorbidities, Respir. Med. 186 (2021) 106528.

[36] H. Jin, I. Vidyanti, P. Di Capua, B. Wu, S. Wu, Predicting Depression among Patients with Diabetes Using Longitudinal Data. Methods Inform. Med., vol. 54 553–559 Preprint at https://doi.org/10.3414/me14-02-0009 (2015).

[37] P. Dworzynski, et al., Nationwide prediction of type 2 diabetes comorbidities, Sci. Rep. 10 (2020) 1776.

[38] X. Wang, J. Eichhorn, I. Haq, A. Baghal, Resting-state brain metabolic fingerprinting clusters (biomarkers) and predictive models for major depression in multiple myeloma patients, PLoS One 16 (2021) e0251026.

[39] A.S. Abdalrada, J. Abawajy, T. Al-Quraishi, S.M.S. Islam, Machine learning models for prediction of co-occurrence of diabetes and cardiovascular diseases: a retrospective cohort study, J. Diabetes Metab. Disord. 21 (2022) 251–261.

[40] V. Oliva, et al., Machine learning prediction of comorbid substance use disorders among people with bipolar disorder, J. Clin. Med. 11 (2022) 3935.

[41] A. Khan, S. Uddin, U. Srinivasan, Chronic disease prediction using administrative data and graph theory: The case of type 2 diabetes, Expert Syst. Appl. 136 (2019) 230–241.

[42] F.S. Roque, et al., Using electronic patient records to discover disease correlations and stratify patient cohorts, PLoS Comput. Biol. 7 (2011) e1002141.

[43] A. Elixhauser, C. Steiner, D.R. Harris, R.M. Coffey, Comorbidity measures for use with administrative data, Med. Care 36 (1998) 8–27.

[44] A. Cassell, et al., The epidemiology of multimorbidity in primary care: a retrospective cohort study, Br. J. Gen. Pract. 68 (2018) e245–e251.

[45] K. Barnett, et al., Epidemiology of multimorbidity and implications for health care, research, and medical education: a cross-sectional study, Lancet 380 (2012) 37–43.

[46] A.M.S. Multimorbidity, a priority for global health research. The Academy of Medical Sciences, Acad. Med. Sci. (2018).

[47] Y. Li, C. Ding, Effects of Sample Size, Sample Accuracy and Environmental Variables on Predictive Performance of MaxEnt Model, Polish J. Ecol. vol. 64 303–312 Preprint at https://doi.org/10.3161/15052249pje2016.64.3.001 (2016).

[48] N.K. Neerchal, H. Lacayo, B.D. Nussbaum, Is a Larger Sample Size Always Better? Am. J. Mathem. Manage. Sci., vol. 28 295–307 Preprint at https://doi.org/10.1080/01966324.2008.10737730 (2008).

[49] UCL Institute of Health Informatics. CALIBER. https://www.ucl.ac.uk/health-informatics/research/caliber (2022).

[50] The SAIL Databank, Swansea University. The HDR UK Phenotype Library. The HDR UK Phenotype Library A Reference Catalogue of Human Diseases https://phenotypes.healthdatagateway.org/ (2023).

[51] J.R. Robinson, W.-Q. Wei, D.M. Roden, J.C. Denny, Defining phenotypes from clinical data to drive genomic research, Annu. Rev. Biomed. Data Sci. 1 (2018) 69–92.

[52] C. Shivade, et al., A review of approaches to identifying patient phenotype cohorts using electronic health records, J. Am. Med. Inform. Assoc. 21 (2014) 221–230.

[53] J. Huang, C.X. Ling, Using AUC and accuracy in evaluating learning algorithms, IEEE Trans. Knowl. Data Eng. 17 (2005) 299–310.

[54] C.X. Ling, J. Huang, H. Zhang, AUC: A Better Measure than Accuracy in Comparing Learning Algorithms. In: Adv. Artific. Intell., 329–341 (Springer Berlin Heidelberg, 2003).

[55] A.J. Bowers, X. Zhou, Receiver Operating Characteristic (ROC) Area Under the Curve (AUC): A Diagnostic Measure for Evaluating the Accuracy of Predictors of Education Outcomes, J. Edu. Students Placed at Risk (JESPAR) 24 (2019) 20–46.

[56] X. Wang, F. Wang, J. Hu, A Multi-task Learning Framework for Joint Disease Risk Prediction and Comorbidity Discovery, In: 2014 22nd International Conference on Pattern Recognition 220–225 (2014).

[57] J.A. Bernstein, L.-P. Boulet, M.E. Wechsler MDMMSc, Asthma, COPD, and Overlap: A Case-Based Overview of Similarities and Differences. (CRC Press, 2018).

[58] K.Y. Ong, P.S.S. Lee, E.S. Lee, Patient-centred and not disease-focused: a review of guidelines and multimorbidity, Singapore Med. J. 61 (2020) 584–590.

[59] M. Costanzo, et al., Global Genetic Networks and the Genotype-to-Phenotype Relationship, Cell 177 (2019) 85–100.

[60] J. Shi, et al., Genotype-Phenotype Association Analysis Reveals New Pathogenic Factors for Osteogenesis Imperfecta Disease, Front. Pharmacol. 10 (2019).

[61] S.N. Payrovnaziri, et al., Explainable artificial intelligence models using real-world electronic health record data: a systematic scoping review, J. Am. Med. Inform. Assoc. 27 (2020) 1173–1185.

[62] P.N. Srinivasu, N. Sandhya, R.H. Jhaveri, R. Raut, From Blackbox to Explainable AI in Healthcare: Existing Tools and Case Studies, Mobile Inform. Syst. 2022 (2022).

[63] M.L. McEntee, B. Gandek, J.E. Ware, Improving multimorbidity measurement using individualized disease-specific quality of life impact assessments: predictive validity of a new comorbidity index, Health Qual. Life Outcomes 20 (2022) 108.

[64] E.B. Cohen, I.K. Gordon, First, do no harm. Ethical and legal issues of artificial intelligence and machine learning in veterinary radiology and radiation oncology, Vet. Radiol. Ultrasound 63 (Suppl 1) (2022) 840–850.

[65] J. Halamka, M. Bydon, P. Cerrato, A. Bhagra, Addressing racial disparities in surgical care with machine learning, NPJ Digit. Med. 5 (2022) 152.