Review Article

DIGITAL HEALTH

DIGITAL HEALTH
Volume 10: 1–12

DOI: 10.1177/20552076241272657
journals.sagepub.com/home/dhj

Sage

The application of explainable artificial intelligence (XAI) in electronic health record research: A scoping review

Jessica Caterson[1], Alexandra Lewin[2,*] and Elizabeth Williamson[2,*]

Abstract

Machine Learning (ML) and Deep Learning (DL) models show potential in surpassing traditional methods including generalised linear models for healthcare predictions, particularly with large, complex datasets. However, low interpretability hinders practical implementation. To address this, Explainable Artificial Intelligence (XAI) methods are proposed, but a comprehensive evaluation of their effectiveness is currently limited. The aim of this scoping review is to critically appraise the application of XAI methods in ML/DL models using Electronic Health Record (EHR) data. In accordance with PRISMA scoping review guidelines, the study searched PUBMED and OVID/MEDLINE (including EMBASE) for publications related to tabular EHR data that employed ML/DL models with XAI. Out of 3220 identified publications, 76 were included. The selected publications published between February 2017 and June 2023, demonstrated an exponential increase over time. Extreme Gradient Boosting and Random Forest models were the most frequently used ML/DL methods, with 51 and 50 publications, respectively. Among XAI methods, Shapley Additive Explanations (SHAP) was predominant in 63 out of 76 publications, followed by partial dependence plots (PDPs) in 11 publications, and Locally Interpretable Model-Agnostic Explanations (LIME) in 8 publications. Despite the growing adoption of XAI methods, their applications varied widely and lacked critical evaluation. This review identifies the increasing use of XAI in tabular EHR research and highlights a deficiency in the reporting of methods and a lack of critical appraisal of validity and robustness. The study emphasises the need for further evaluation of XAI methods and underscores the importance of cautious implementation and interpretation in healthcare settings.

Keywords

Artificial intelligence, digital, digital health, eHealth, electronic, health, health informatics, machine learning, systematic reviews

Submission date: 3 April 2024; Acceptance date: 9 July 2024

Introduction

The 'Black Box' problem

Artificial intelligence (AI) in healthcare is anticipated to transform the sector.[1–7] One example is machine learning (ML) and deep learning (DL) models for predictive analytics using tabular electronic health record (EHR) data.[8] EHR data are a wealthy resource, providing extensive information which can be applied for many beneficial applications including predicting 30-day mortality,[9] identifying risk groups for future diseases,[10,11] and estimating hospital length of stay.[12]

However, a limitation of ML/DL applications in this context is that how or why the predictions are made from a set of covariable, or feature, values is generally unclear.[1,13–17] Such explanations are necessary in order to justify, control, improve and discover new information from any model.[14] Justification is important to ensure patient and clinician trust in application of these models in health,[15,18,19] and to comply with ethical and legal

[1]Imperial College London, London, UK
[2]London School of Hygiene and Tropical Medicine, Bloomsbury, UK
[*]Joint last authors.

Corresponding author:
Elizabeth Williamson, London School of Hygiene and Tropical Medicine, Keppel St, London WC1E 7HT, UK.
Email: elizabeth.williamson@lshtm.ac.uk

expectations. Any AI model implemented in health settings will need to gain trust from the patients it is used on, and the clinicians and organisations that use it. Failure to do so may risk the under-utilisation of AI in areas where it can be of significant benefit.[20] From an ethical perspective, humans should be treated as such, not as objects, and thus 'human in the loop' approaches are preferrable.[21] Legally, there are an increasing number of rules and regulations which set expectations around the explainability of AI.[21] For example, the General Data Protection Regulation (GDPR) stipulates the right to an explanation and to contest any decision relating to automated processing and a person's health.[22,23] GDPR also highlights the importance of informed consent in the context of automated decision making,[23] a sentiment echoed in other legal documents such as the European Chart of Fundament Rights.[24] Control and improving models are essential to ensure any predictive model is not biased, and protect the model from adversarial attacks.[14,19,25–27] If predictions can be explained, then the underlying mechanisms for disease could be discovered – however, extra care must be taken in separating an association from a causal pathway.[18,28–30]

This lack of interpretability of complex ML/DL methods is referred to as the 'black box problem'.[1,13] This contrasts to methods such as generalised linear models (GLMs), decision trees (DTs) and rule-based methods, which are 'intrinsically interpretable'. That is they generate easily interpretable outputs which are intuitively explainable to humans.[18]

Explainable artificial intelligence (XAI)

A proposed solution for interpretation of 'black box' models is explainable artificial intelligence (XAI).[14,19] XAI encompasses a range of interpretation methods which can be applied to black box models to generate interpretable explanations for their predictions. Approaches include 'model-specific' methods limited to one or some ML/DL models, and 'model-agnostic' applicable to any ML/DL model.[31] Methods may also be 'intrinsic', i.e., they arise from the model itself (just as coefficient weights are produced from GLMs, and tree visualisations from decision trees), or 'post-hoc', requiring further analysis for the explanation.[31] There are also 'transparent models', which are designed to retain the high-performing predictive capabilities of ML/DL, but have the interpretability of simpler models.[15,18,19] They include generalised additive models (GAMs)[32] and Bayesian Rule Lists (BRLs).[33,34] However, their predictive performance is regarded as inferior to standard ML/DL methods, and thus they are not seen as a viable alternative at present.[15]

Among XAI methods, model-agnostic post-hoc approaches are preferrable, preserving flexibility for model selection and interpretation comparison.[31] Post-hoc methods also do not require the underlying model to be altered, and so can be developed without access to the original model training data. These approaches can be further sub-divided according to the nature of the explanation provided into: (1) individual explanations (2) global explanations and (3) feature/outcome relationships.

Individual explanations. Individual explanations provide information for why a specific prediction was made e.g., why a particular patient with certain characteristics is predicted to developed breast cancer within the next 5 years. This is advantageous when explaining black box model predictions, because many of the modelling methods employ multi-dimensional, complex, non-linear explanations which are difficult to generalise across all individuals.[31,35] By focussing explanations on one individual of interest, individualised XAI methods offer an explanation that is locally representative.[35]

XAI methods for individual explanations include Local Individualised Model-Agnostic Explanations (LIME),[35] Shapley Values,[36] SHapley Additive exPlanations (SHAP)[37] and 'Counterfactual' explanations.[31]

The LIME method creates a locally representative explanation by building a linear model around an individual of interested and surrounding data points (e.g., the patient of interest and patients with similar characteristics) weighted inversely by their distance from the point of interest.[35] The prediction is explained by coefficient weights derived from the linear model for each feature. The weight of the data points around the individual of interest is defined heuristically.[35] Thus, explanations are vulnerable to changes in the weighting used.

Shapley values are derived from economic game theory, where they calculate the 'pay-out' to each contributor working in a coalition to generate profit.[36] In XAI, the contributors are the features in the model, and the 'profit' is the predicted outcome. Shapley values represent the contribution of each feature to the final prediction; the larger the contribution, the larger the Shapley value 'pay-out'.

SHAP values are derived from Shapley values. However, they are constructed such that the sum of the SHAP values for all of the features for a single prediction is equal to the final prediction. This is an example of an additive attributive model.[37] In doing so, SHAP values can be aggregated across predictions to give global and feature/outcome relationship explanations. Calculation of SHAP values is computationally intensive, thus in practice, a user-defined sample of data is usually taken, and the LIME method is used to calculate the SHAP values.[31,37]

Counterfactual explanations utilise an interpretation approach commonly used by humans where something is explained by describing what would need to happen for the prediction to change.[31] In the context of predictive models, this involves illustrating how the predicted outcome changes given changes in input features/variables.[38] This is favourable to allow non-technical audiences

to conceptualise unknown modelling processes by associating change of input with output.[39] However, they do not necessarily equate to causality, which is a clear risk of their application in healthcare settings.

Global explanations. Global explanations provide an overview of how features contribute to predicted outcomes.[31] Examples include global surrogate models and feature importance.

Global surrogate models train an intrinsically interpretable model on the features from the original data and predicted outcome from the black box model, to generate an intrinsically interpretable model which – hopefully – closely replicates the predictions from the more complex black box model.[31] The explanation from the intrinsically interpretable model can then explain how the predictions are generated from the original black box model. It is also important to ensure that the interpretable model is a reasonable proxy for the black box model. Molnar (2023) suggests this can be done by calculating the R^2 between the predictions from the black box model and those from the surrogate model.[31]

Feature Importance describes multiple methods which rank features used to train a model in order of some metric of 'importance'. The earliest feature importance was designed for Random Forests.[40] This method retrains the model with each feature's values consecutively shuffled, and compares the difference in 'impurity' (i.e., the lack of discrimination between the true and predicted outcome) between each model and the original model. The greater the difference in impurity, the more important that feature. Later, this method could apply to any model type by comparing prediction error.[41] Feature importance is a useful aid for feature selection, but is less reliable for model interpretation as it depends on the scale of the feature's values. In addition, it may be influenced by unrealistic permuted data values.[31] Features may also be regarded as less important if correlated with each other.[31]

Contrastingly, SHAP feature importance takes the absolute SHAP value for each feature for all individuals in the data set and averages the value.[37] It then ranks features in importance of average absolute SHAP value. This is a truly 'post-hoc' and 'model-agnostic' approach and, rather than describing error, quantifies overall contributions of each feature to predicted outcomes.

Feature/outcome relationships. Feature/Outcome relationships specify how changing a feature value changes the prediction, including non-linear and multi-dimensional associations.[31] Relationship explanations can also show how different features interact with each other by colouring plots according to an additional feature's value. Relationship explanations are often displayed graphically as a feature value against prediction, and thus are typically limited to showing 1-2 features at a time due to the limits of human dimensional perception.[42] Examples of methods include partial dependence plots (PDPs), individual expectation curves (ICE) and accumulated local effects (ALE) plots.

Partial dependence plots show how a prediction changes over a range of feature values across their marginal distribution. That is, overall values of other features in the dataset. The relationship is calculated using the partial dependence function, which for a given feature, takes each individual's data and calculates the predicted outcome for that feature value with the remaining values fixed.[42–44] The final relationship is an average of how the prediction changes over all individuals. A risk of using these plots is over-interpreting relationships for specific feature values which are very rare or improbable in the dataset. This can be addressed by showing density plots, which show how the feature values are distributed. ICE plots are very similar, except they display each individual as a separate curve on the plot.[45] This helps to show heterogeneity in changes in prediction between individuals.

ALE plots use the conditional instead of marginal distribution of the feature being changed.[31] This avoids improbable or impossible combinations of values,[31] for example a person of height 2 m, who weighs 40 kg, or a person aged 50 who has smoked for 45 years.

It is also possible to display SHAP dependence plots.[37] These plots display the relationship of feature value to its SHAP value across a dataset.

XAI in healthcare research

Research using tabular EHR data are increasingly adopting XAI methods. A scoping review conducted by Payrovnaziri et al. (2020) from 2009 to 2019 identified only 42 publications using XAI, with only 5 employing post-hoc and model-agnostic methods.[46] LIME appeared in two publications,[47,48] and a global surrogate model in one.[49] Unique methods included a probability calculation for feature importance by Eck et al.,[50] and automated/manual rule pruning by Luo et al.[51] A later survey conducted by Di Martino and Delmastro (2023) up to 2021 reported 71 publications using post-hoc model-agnostic XAI in tabular and time-series EHR research, with 90% published from 2020 onwards.[52] SHAP was the most frequently used interpretation method in this survey.[52]

Study aims/objectives

The aim of this scoping review is to provide an up-to-date overview into the use of post-hoc model-agnostic interpretation methods for ML and DL in EHR tabular research. It also aims to characterise and critically appraise the use of frequently applied methods in practice.

Methods

The scoping review was conducted in accordance with PRISMA guidelines for scoping reviews.[53] Table 1 provides the inclusion and exclusion criteria. Two major scientific literature databases, PUBMED and OVID/MEDLINE (including EMBASE) were searched for publications.

The review comprised three phases: an initial broad search, title and abstract review, and full-text review. The initial search was conducted on 1st May 2023 and an updated search was completed on 27th June 2023, using the search terms provided in Figure 1.

All publications including original articles, reviews, and letters were included in the initial search, published any year, and available in English. The purpose of starting with a broad search field was to capture the wide range of terminology and publication types across computer science and medical science research. The online systematic review platform Rayyan was used to carry out publication screening.[54] In the two subsequent phases of screening, title and abstract, and full text, the exclusion and inclusion criteria in Table 2 were applied. Publications were included if they used post-hoc model-agnostic XAI to explain ML models applied to tabular EHR data in healthcare. Decision to exclude articles was based on three criteria: out of scope (for example, no XAI, or only intrinsic or model-specific XAI e.g., methods specific to deep learning, neural networks), wrong data type (for example imaging, genetic data, or natural language processing), or if the full text was unavailable. A final search of grey literature from included publication citations and the Google search engine identified additional relevant articles.

Title, publication type, year of publication, author, journal, medical sub-specialty and study aim were recorded for each selected article. In addition, ML models used (including final selected model), metrics for model selection, and interpretation method(s). The nature of interpretation was analysed under three domains: (1) Global Feature Importance, (2) Feature/Outcome Relationships and (3) Individual Explanations. Data were analysed and summarised using R (v4.2.2).

Table 1. Inclusion and exclusion criteria applied for the scoping review literature search.

Inclusion criteria	Exclusion criteria
Publication relating to the use of interpretation methods (XAI) to explain machine learning models in a healthcare setting.	Outside scope (no XAI, intrinsic or model-specific XAI only)
Post-hoc model-agnostic XAI used	Wrong Data Type (Related to genetics, imaging, natural language processing, or sensor data)
Tabular electronic healthcare data research	Full Text Unavailable
Original articles, reviews, letters	
Text available in English	

XAI = explainable AI.

Results

After de-duplication, 3220 publications were identified from the search criteria. Of these, 132 were selected for full-text review. There were 2792 publications excluded for being outside the scope of the review and a further 296 related to clinical imaging or genetics. Of the 132 selected for full-text review, 75 were selected for final inclusion. An additional paper was included from the grey literature search. Figure 2 summarises the selection process.

Publication characteristics

All included publications were original peer-reviewed research articles. The articles were published between February 2017 to June 2023, increasing substantially in this time period (Figure 3a).

Identified studies aimed to improve predictive performance for classification of diseases, identifying complications and estimating length of stay using EHR data. Common domains were medicine (24 publications),[48,55–77] COVID-19 (10 publications),[78–87] Psychiatry (7 publications)[88–94] and surgery (7 publications).[95–101] Most publications emphasised interpretation methods for causal reasoning, especially for COVID-19, where knowledge of the disease was limited at the time of publications.[86,87] Justification of the model's predictive performance was also a commonly cited motivation for using interpretation methods.[56,102–104]

Machine learning models

Of the 76 publications, most trialled multiple ML/DL methods to get the best predictive performance. In the model development phase, EXtreme Gradient Boosting trees (XGB) were the most used in 67% (51/76) publications, followed by Random forest (RF) in 66% (50/76) of studies and Generalised linear models like logistic regression (LR) (Figure 3b). XGB remained the top choice for the final models, with RF also frequently selected (Figure 3b).

A range of performance metrics were evaluated for the selection of the final ML model (Table 2). Area under the receiver operating curve (AUROC) was the most frequently reported (70/76). F1-score, accuracy, recall (sensitivity), precision (i.e., positive predictive value, PPV) and specificity were also all assessed in over one-third of publications (Table 2).

(((machine learning) OR (ML)) OR ((artificial intelligence) OR (AI)) OR ((deep learning) OR (DL)) AND ((electronic health* record) OR (EHR)) AND ((interpret*) OR (explain*) OR (XAI))

Figure 1. Search terms for PubMed and OVID/MEDLINE (including EMBASE) database.

Table 2. Frequency of metrics for evaluating and selecting the final ML model.

Metric	No. of publications (%)
AUROC	70 (92)
F-score	32 (42)
Accuracy	25 (33)
Recall (sensitivity)	47 (62)
Precision (PPV)	37 (49)
Specificity	25 (33)
NPV	10 (13)

AUROC = area under receiver operater curve, PPV = positive predictive value, NPV = negative predictive value.

Interpretation method selection

Overall, 16 different interpretation methods were used (Table 3). SHAP was the most frequently used interpretation method, reported in 83% (63/76) of publications. LIME, PDP and surrogate models were used in 9 (12%), 8 (11%) and 7 (9%) of publications, respectively. Eleven publications included feature importance. Interpretation methods were used to evaluate only the final selected ML model(s) in all bar two publications, which used an interpretation method to compare how different models predicted outcomes.[55,84]

Interpretation method application

Individual explanations. SHAP values were the most frequently used method for individual explanations, presented in 30 publications.[55,56,60,67,68,72–75,77,83,86,89,91,93,98–100,103–114] Individual explanations using the SHAP method were displayed visually with force or waterfall plots.[111] These plots show a series of arrows pointing right (positive SHAP value) or left (negative SHAP value), which together combine to reach the final predicted outcome.

LIME was used in eight publications.[48,55,73,84,97,102,115,116] Sun et al. (2022) compared SHAP individual explanations with LIME, and inferred that their similar interpretations supported the 'stability' of the interpretations, which may be expected given SHAP values are calculated from the LIME method in practice.[73] LIME was commonly used to contrast positive and negative predictions.[48,73,115,116] Alternatively, Kibria et al. (2022) used LIME to compare outputs for one prediction across models to explain why different black box models had different predictive performances.[55]

Amongst all the publications that used SHAP or LIME, only one publication reported the heuristically defined hyper-parameters.[48]

Other, less frequently used individual explanations included counterfactual explanations. There was only one identified use of this method by Banerjee et al. (2021) to predict mortality in severe mental illness.[92] However, rather than using this method causally as is the intended purpose, they reported a spurious result which they felt was indicative of bias in the dataset.

Global explanation methods. SHAP importance was the most frequently reported metric. In its simplest form, SHAP importance was represented as a bar plot.[60,63,67,69,70,73,75,77,80–82,86–90,98,99,101,104,105,110,113,117–125] However, another commonly used figure was the 'violin plot', which shows the SHAP value for each instance in the data as a dot for each feature.[55–59,62,66,72,74,76,78,79,83,84,91,93–96,100,103,106,109,112,114,126–128] Features are ordered from top to bottom from the most to least important. Each dot is coloured to represent if the feature value is high or low. In doing so, this plot not only shows the ranked order of importance of features, but how feature values may influence the SHAP value for each feature. Feature importance was used in 11 publications.[48,55,57,60,66,73,99,108,112,113,119] It was typically used as simple summary of a model, followed by other explanations.[48,55,57,60,66,73,99,101,108,111–113,119,124] In some cases, feature importance was compared with SHAP importance.[60,73,99,113,119] The order of features according to permutation feature importance *versus* SHAP importance often differed, but was little commented on. This could have implications in later work, which may only select some features for further investigation based on one or other important metric.

Global surrogate models were used in six publications.[49,82,114,116,123,129] This method was generally used to justify an underlying black box model, such as describing

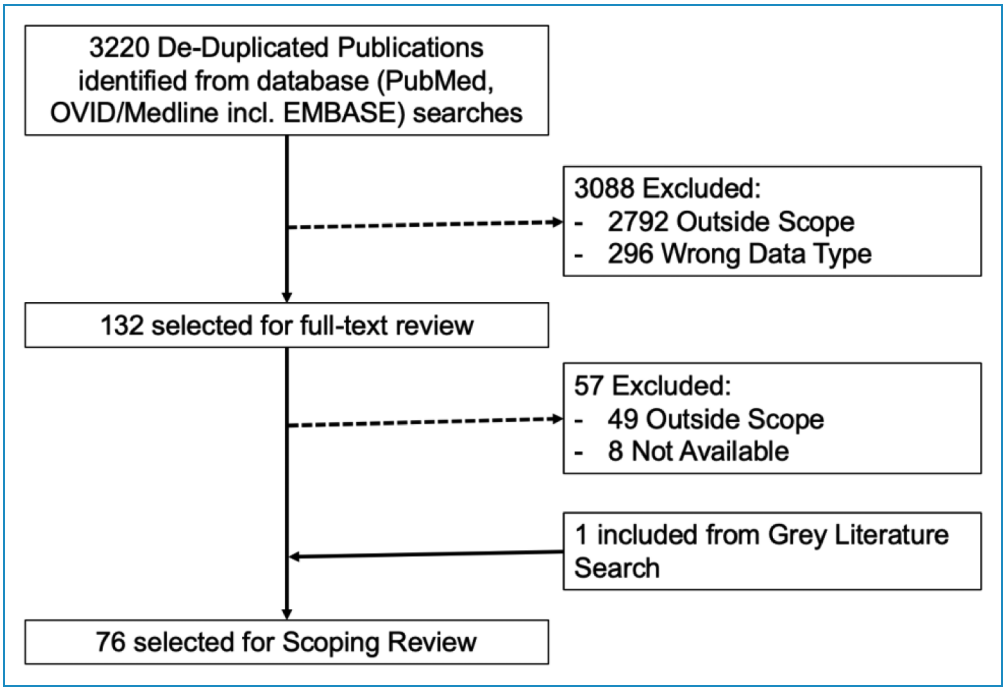

Figure 2. PRISMA scoping review flow chart.

predictive pathways. Vyas et al. (2022), used a surrogate model to describe predictions for the diagnosis and severity of dementia.[116] They explained their final (black box) model, a random forest, using a simple decision tree. From the decision tree, they identified groups at very high or low risk of developing dementia, and stated, for example, 'if a person fails to identify all three animals [in a cognitive assessment questionnaire], they have a 95% chance of developing dementia'.[116] Only one (Zhang et al. 2022[123]) publication quantified how well the surrogate model approximated the black box model. Zhang et al. (2022), reported R^2 values of 0.68 and 0.57 for their support vector machine (SVM) and RF models, i.e., the proportion of variance explained between predictions from the black box model and surrogate model were 68% and 57%, respectively.[123] These values implied there was a difference between the surrogate and underlying black box model, and may invalidate the surrogate model as a means to justify or provide an explanation for this prediction.

Feature/outcome relationships. Partial Dependence Plots were used in 15 publications.[49,61,66,71,73,79,100,101,108,109,111,121,125] Relationships between features and outcomes were shown over a wide range of feature values, however few publications also showed the density of data at feature values from which the calculations were made.[37,79,109]

In two publications, individual conditional expectation (ICE) curves were also shown.[73,108] In the paper by Sun et al. (2022), this was helpful in highlighting heterogeneity in feature/outcome relationships for some features, which can be hidden by showing only the average partial dependence alone.[73]

Sun et al. (2022) and Qiu et al. (2022) also presented SHAP dependence plots with PDPs to compare relationships.[73, 109] The plots were generally similar in both studies, which may imply the techniques are robust. Interactions between features were shown using SHAP dependence plots in eight publications.[57,59,93,94,96,98,112,114]

Discussion

This scoping review shows a significant rise in the interest and application of XAI methods within the last 5 years in EHR tabular research. In contrast to the 2020 review by Payrovnaziri,[46] this review reports a 10-fold increase in post-hoc model-agnostic publications. This aligns with the study by Di Martino & Delmastro (2023) which reported a substantial rise in XAI publications up to 2021.[52] Among the 76 identified publications, diverse method choices and motivations were evident. Some focussed on simple global feature importance metrics, whilst others only have individual explanations, and some gave comprehensive interpretations of global importance, feature/outcome relationships and individual examples.

The need for XAI was supported by the high rate of selection of black box models as the final model (e.g., Random Forest, XGBoost, Neural Networks), although there is likely a high selection bias towards these methods as simpler interpretable methods were less likely to need

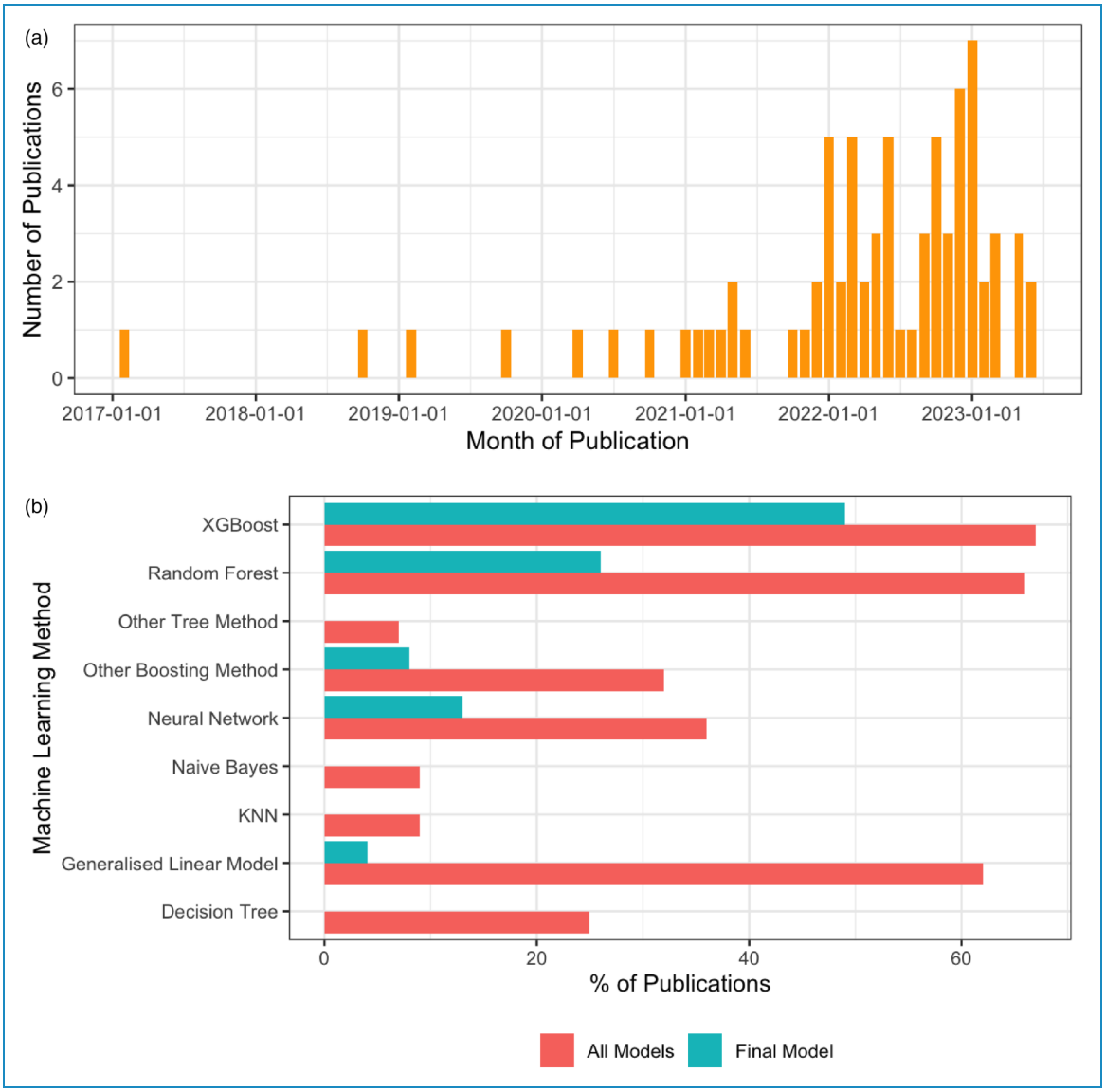

Figure 3. (a) Included publications over calendar time (b) Machine learning methods selected for testing and the final model as a % of publications.

XAI in the first place. AUROC was overall the most favoured metric, whereas other metrics were seen in a third or less of publications. AUROC may have been favoured for its interpretation in the context of imbalanced datasets, which are commonplace in healthcare data. The range of other metrics selected may represent the diverse aims of the models developed which may favour a specific goal e.g., high sensitivity versus high specificity.

SHAP was the most popular method, not reported by Payrovnaziri et al. (2020),[46] but by Di Martino & Delmastro (2023).[52] This is likely because the SHAP method was introduced in 2017,[37] and so only translated to healthcare research post-2019. Its widespread use suggests a perceived superior performance versus other methods, but authors seldom explicitly justify this choice.[52] However, what is notable about SHAP compared with the other methods is that it offers interpretation approaches across all three core domains discussed. It also is more theoretically robust than other techniques.[31,37,52] Thus, using the SHAP method enables authors to gain in-depth insight into their black box models, with a strong theoretical backing,[31] without needing to use multiple methods.

The interpretation methods lack critical appraisal, with vulnerabilities such as susceptibility to adversarial attacks, as seen in instances like concealing ethnicity in the COMPAS dataset.[130] Potential consequences, such as influencing predictions in healthcare decision-making, remain largely unexplored and unacknowledged in publications. Moreover, the dependence on heuristically defined hyperparameter values is evident,[35,37] with limited reporting

Table 3. Methods for interpretation of machine learning and the number (percentage) of publications.

Method	Interpretation domain	No. of publications (%)
SHAP	Global, relationships, individual	63 (83)
PDP	Relationships	15 (20)
Feature importance	Global	11 (14)
LIME	Individual	8 (11)
Surrogate Model	Global	6 (8)
Shapley	Individual	3 (4)
ICE	Relationship	3 (4)
Rule-based	Global	1 (1)
Counterfactual	Individual	1 (1)
Other	–	7 (9)

and understanding of vulnerability by authors.[48] Evaluation of global surrogate models was also lacking, and in the one instance where it was performed, there was poor approximation between the black box and surrogate model.[123] Evaluating similarity should be a requirement when using this method. Partial dependence plots were shown over wide feature value spaces, even though some feature values may be very rare or impossible.[31] Greater awareness of the limitations of these interpretation methods is important to their appropriate application in the future.[52]

This scoping review is limited to tabular electronic health record research, and thus does not provide an overview of the application of interpretation methods to other healthcare domains such as genetics, imaging, NLP and time-series analysis. XAI in these domains differs from tabular EHR research, due to the nature of the data involved,[52] and thus the applicability of XAI methods. This review also does not report model-specific methods, methods limited to subdomains of AI such as deep neural networks, or transparent models, and thus may not include other widely used methods limited to these groups. However, these methods are arguably inferior to model-agnostic approaches, given they limit the flexibility in initial model selection.[14,15,31,52]

Conclusion

Overall, this scoping review reinforces the recent surge in the application of post-hoc interpretation methods for addressing the black box problem in health care. There is significant heterogeneity in the choice and application of these methods. There is also a lack of evaluation of XAI methodologies. Further research is required to properly assess these methods for their intended purpose and ensure that they are appropriate, robust and stand up to the demands of explanations expected within healthcare.

Contributorship: JC defined the search question, AL and EW helped refine the question. JC conducted the literature review, conducted the analysis and wrote the first draft. AL and EW reviewed the analysis and edited the final draft for approval. All authors reviewed and edited the manuscript and approved the final version.

Declaration of conflicting interests: The authors declared no potential conflicts of interest with respect to the research, authorship and/or publication of this article.

Ethical approval: The ethics committee of the London School of Hygiene and Tropical Medicine approved this study (REF: 28983-01).

Funding: The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research was funded in part by the Wellcome Trust [Senior Research Fellowship 224485/Z/21/z]. For the purposes of open access, the author has applied a CC BY copyright licence to any Author Accepted Manuscript version arising from this submission.

Reporting guidelines: PRISMA Checklist is included as supplementary material.

Guarantor: JC

ORCID iD: Jessica Caterson https://orcid.org/0000-0002-4773-1797

Supplemental material: Supplemental material for this article is available online.

References

1. Castelvecchi D. Can we open the black box of AI? *Nature* 2016; 538: 20–23.
2. Cohen IG, Amarasingham R, Shah A, et al. The legal and ethical concerns that arise from using complex predictive analytics in health care. *Health Aff (Millwood)* 2014; 33: 1139–1147.
3. Sheu R-K and Pardeshi MS. A survey on medical explainable AI (XAI): recent progress, explainability approach, human interaction and scoring system. *Sensors (Basel)* 2022; 22: 8068.
4. Esmaeilzadeh P. Use of AI-based tools for healthcare purposes: a survey study from consumers' perspectives. *BMC Med Inform Decis Mak* 2020; 20: 170.

5. Houben S, Abrecht S, Akila M, et al. Inspect, understand, overcome: a survey of practical methods for AI safety. In: *Deep neural networks and data for automated driving*. Cham: Springer International Publishing, 2021, pp.3–78.
6. Shickel B, Tighe PJ, Bihorac A, et al. Deep EHR: a survey of recent advances in deep learning techniques for electronic health record (EHR) analysis. *IEEE J Biomed Heal Informatics* 2018; 22: 1589–1604.
7. Pettit RW, Fullem R, Cheng C, et al. Artificial intelligence, machine learning, and deep learning for clinical outcome prediction. *Emerg Top Life Sci* 2021; 5: 729–745.
8. James G, Witten D, Hastie T, et al. *An Introduction to Statistical Learning*. New York, NY: Springer US, 2021. DOI: 10.1007/978-1-0716-1418-1.
9. Choi MH, Kim D, Choi EJ, et al. Mortality prediction of patients in intensive care units using machine learning algorithms based on electronic health records. *Sci Rep* 2022; 12: 7180.
10. Kilic A. Artificial intelligence and machine learning in cardiovascular health care. *Ann Thorac Surg* 2020; 109: 1323–1329.
11. Ambale-Venkatesh B, Yang X, Wu CO, et al. Cardiovascular event prediction by machine learning: the multi-ethnic study of atherosclerosis. *Circ Res* 2017; 121: 1092–1101.
12. Bacchi S, Tan Y, Oakden-Rayner L, et al. Machine learning in the prediction of medical inpatient length of stay. *Intern Med J* 2022; 52: 176–185.
13. Burrell J. How the machine 'thinks': understanding opacity in machine learning algorithms. *Big Data Soc* 2016; 3: 205395171562251.
14. Adadi A and Berrada M. Peeking inside the Black-Box: a survey on explainable artificial intelligence (XAI). *IEEE Access* 2018; 6: 52138–52160.
15. Guidotti R, Monreale A, Ruggieri S, et al. A survey of methods for explaining Black Box models. *ACM Comput Surv* 2019; 51: 1–42.
16. Suman RR, Mall R, Sukumaran S, et al. Extracting state models for Black-Box software components. *J Object Technol* 2010; 9: 79.
17. Kelly CJ, Karthikesalingam A, Suleyman M, et al. Key challenges for delivering clinical impact with artificial intelligence. *BMC Med* 2019; 17: 195.
18. Lipton ZC. The Mythos of Model Interpretability, http://arxiv.org/abs/1606.03490 (2016, accessed 23 June 2023).
19. Abdul A, Vermeulen J, Wang D, et al. Trends and trajectories for explainable, accountable and intelligible systems. In: *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, New York, NY, USA: ACM, pp.1–18.
20. Pagallo U, O'Sullivan S, Nevejans N, et al. The underuse of AI in the health sector: opportunity costs, success stories, risks and recommendations. *Health Technol (Berl)* 2024; 14: 1–14.
21. Stöger K, Schneeberger D and Holzinger A. Medical artificial intelligence: the European legal perspective. *Commun ACM* 2021; 64: 34–36.
22. *What is automated individual decision-making and profiling?*, https://ico.org.uk/media/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/automated-decision-making-and-profiling/what-is-automated-individual-decision-making-and-profiling-1-1.pdf (2018, accessed 27 November 2022).
23. Art. 22 GDPR – Automated individual decision-making, including profiling – GDPR.eu, https://gdpr.eu/article-22-automated-individual-decision-making/ (accessed 27 November 2022).
24. Article 3 – Right to integrity of the person | European Union Agency for Fundamental Rights, https://fra.europa.eu/en/eu-charter/article/3-right-integrity-person (accessed 29 May 2024).
25. Kulesza T, Wong W-K, Stumpf S, et al. Fixing the program my computer learned. In: *Proceedings of the 14th International Conference on Intelligent User Interfaces*, New York, NY, USA: ACM, pp.187–196.
26. Kulesza T, Stumpf S, Burnett M, et al. Tell me more? In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, New York, NY, USA: ACM, pp.1–10.
27. Stumpf S, Rajaram V, Li L, et al. Interacting meaningfully with machine learning systems: three experiments. *Int J Hum Comput Stud* 2009; 67: 639–662.
28. Liu C, Rani P and Sarkar N. An empirical study of machine learning techniques for affect recognition in human-robot interaction. In: *2005 IEEE/RSJ International Conference on Intelligent Robots and Systems*, IEEE, pp.2662–2667.
29. Pearl J. *Causality: models, reasoning, and inference*. Cambridge, UK: Cambridge University Press, 2000.
30. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat Mach Intell* 2019; 1: 206–215.
31. Molnar C. *Interpretable machine learning: a guide for making Black Box Models interpretable*, https://www.amazon.co.uk/Interpretable-Machine-Learning-Christoph-Molnar/dp/0244768528 (accessed 26 June 2023).
32. Caruana R, Lou Y, Gehrke J, et al. Intelligible models for HealthCare. In: *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, NY, USA, pp.1721–1730: ACM.
33. Letham B and Rudin C. Building Interpretable Classifiers with Rules using Bayesian Analysis, https://www.semanticscholar.org/paper/Building-Interpretable-Classifiers-with-Rules-using-Letham-Rudin/684cad92335f9690d56a4c27d47b08c408b05d10 (2012, accessed 26 June 2023).
34. Souillard-Mandar W, Davis R, Rudin C, et al. Interpretable Machine Learning Models for the Digital Clock Drawing Test, http://arxiv.org/abs/1606.07163 (2016, accessed 26 June 2023).
35. Ribeiro MT, Singh S and Guestrin C. 'Why Should I Trust You?' Explaining the Predictions of Any Classifier, http://arxiv.org/abs/1602.04938 (2016, accessed 26 June 2023).
36. Shapley LS. 17. A value for n-person games. In: *Contributions to the theory of games (AM-28), volume II*. Princeton, New Jersey, USA: Princeton University Press, 1953, pp.307–318.
37. Lundberg S and Lee S-I. A Unified Approach to Interpreting Model Predictions, http://arxiv.org/abs/1705.07874 (2017, accessed 26 June 2023).

38. Roese NJ. Counterfactual thinking. *Psychol Bull* 1997; 121: 133–148.
39. Del Ser J, Barredo-Arrieta A, Díaz-Rodríguez N, et al. On generating trustworthy counterfactual explanations. *Inf Sci (Ny)* 2024; 655: 119898.
40. Breiman L. Random forests. *Mach Learn* 2001; 45: 5–32.
41. Fisher A, Rudin C and Dominici F. All models are wrong, but many are useful: learning a variable's importance by studying an entire class of prediction models simultaneously. *J Mach Learn Res* 2019; 20: 177.
42. Friedman JH. Greedy function approximation: a gradient boosting machine. *Ann Stat* 2001; 29: 1189–1232.
43. Hastie T, Tibshirani R and Friedman J. *Springer Series in Statistics The Elements of Statistical Learning Data Mining, Inference, and Prediction*, https://hastie.su.domains/Papers/ESLII.pdf (accessed 26 June 2023).
44. Greenwell BM. *PDP: an R Package for Constructing Partial Dependence Plots*, https://github.com/bgreenwell/pdp/issues. (accessed 26 June 2023).
45. Goldstein A, Kapelner A, Bleich J, et al. Peeking inside the Black Box: visualizing statistical learning with plots of individual conditional expectation. *J Comput Graph Stat* 2015; 24: 44–65.
46. Payrovnaziri SN, Chen Z, Rengifo-Moreno P, et al. Explainable artificial intelligence models using real-world electronic health record data: a systematic scoping review. *J Am Med Inform Assoc* 2020; 27: 1173–1185.
47. Ghafouri-Fard S, Taheri M, Omrani MD, et al. Application of single-nucleotide polymorphisms in the diagnosis of autism spectrum disorders: a preliminary study with artificial neural networks. *J Mol Neurosci* 2019; 68: 515–521.
48. Pan L, Liu G, Mao X, et al. Development of prediction models using machine learning algorithms for girls with suspected central precocious puberty: retrospective study. *JMIR Med Informatics* 2019; 7: e11728.
49. Che Z, Purushotham S, Khemani R, et al. Interpretable deep models for ICU outcome prediction. *AMIA. Annu Symp Proc* 2016; 2016: 371–380.
50. Eck A, Zintgraf LM, de Groot EFJ, et al. Interpretation of microbiota-based diagnostics by explaining individual classifier decisions. *BMC Bioinf* 2017; 18: 441.
51. Luo G. Automatically explaining machine learning prediction results: a demonstration on type 2 diabetes risk prediction. *Heal Inf Sci Syst* 2016; 4: 2.
52. Di Martino F and Delmastro F. Explainable AI for clinical and remote health applications: a survey on tabular and time series data. *Artif Intell Rev* 2023; 56: 5261–5315.
53. Tricco AC, Lillie E, Zarin W, et al. PRISMA extension for scoping reviews (PRISMA-ScR): checklist and explanation. *Ann Intern Med* 2018; 169: 467–473.
54. Ouzzani M, Hammady H, Fedorowicz Z, et al. Rayyan – a web and mobile app for systematic reviews. *Syst Rev* 2016; 5: 210.
55. Kibria HB, Nahiduzzaman M, Goni MOF, et al. An ensemble approach for the prediction of diabetes Mellitus using a soft voting classifier with an explainable AI. *Sensors (Basel)* 2022; 22: 7268.
56. Lv J, Zhang M, Fu Y, et al. An interpretable machine learning approach for predicting 30-day readmission after stroke. *Int J Med Inform* 2023; 174: 105050.
57. Lu S, Chen R, Wei W, et al. Understanding heart failure patients EHR clinical features via SHAP interpretation of tree-based machine learning model predictions. *AMIA Annu Symp Proc* 2021; 2021: 813–822.
58. Momenzadeh A, Shamsa A and Meyer J. Bias or biology? Importance of model interpretation in machine learning studies from electronic health records. *JAMIA open* 2022; 5: ooac063.
59. Chen B, Ruan L, Yang L, et al. Machine learning improves risk stratification of coronary heart disease and stroke. *Ann Transl Med* 2022; 10: 1156.
60. Chen W, Li X, Ma L, et al. Enhancing robustness of machine learning integration with routine laboratory blood tests to predict inpatient mortality after intracerebral hemorrhage. *Front Neurol* 2021; 12: 790682.
61. Kim K, Yang H, Yi J, et al. Real-time clinical decision support based on recurrent neural networks for in-hospital acute kidney injury: external validation and model interpretation. *J Med Internet Res* 2021; 23: e24120.
62. Schallmoser S, Zueger T, Kraus M, et al. Machine learning for predicting micro- and macrovascular complications in individuals with prediabetes or diabetes: retrospective cohort study. *J Med Internet Res* 2023; 25: e42181.
63. Yu S, Hofford M, Lai A, et al. Respiratory support status from EHR data for adult population: classification, heuristics, and usage in predictive modeling. *J Am Med Inform Assoc* 2022; 29: 813–821.
64. Shetty M, Kunal S, Girish M, et al. Machine learning based model for risk prediction after ST-elevation myocardial infarction: insights from the North India ST elevation myocardial infarction (NORIN-STEMI) registry. *Int J Cardiol* 2022; 362: 6–13.
65. Wu J, Liu C, Xie L, et al. Early prediction of moderate-to-severe condition of inhalation-induced acute respiratory distress syndrome via interpretable machine learning. *BMC Pulm Med* 2022; 22: 193.
66. Wu R, Shu Z, Zou F, et al. Identifying myoglobin as a mediator of diabetic kidney disease: a machine learning-based cross-sectional study. *Sci Rep* 2022; 12: 21411.
67. Huang J, Jin W, Duan X, et al. Twenty-eight-day in-hospital mortality prediction for elderly patients with ischemic stroke in the intensive care unit: interpretable machine learning models. *Front Public Health* 2022; 10: 1086339.
68. Levy J, Lima J, Miller M, et al. Machine learning approaches for hospital acquired pressure injuries: a retrospective study of electronic medical records. *Front Med Technol* 2022; 4: 926667.
69. Daluwatte C, Dvaretskaya M, Ekhtiari S, et al. Development of an algorithm for finding pertussis episodes in a population-based electronic health record database. *Hum Vaccin Immunother* 2023; 19: 2209455.
70. Emdad F, Tian S, Nandy E, et al. Towards interpretable multimodal predictive models for early mortality prediction of hemorrhagic stroke patients. *AMIA Jt Summits Transl Sci Proc* 2023; 2023: 128–137.
71. Miran S, Nelson S and Zeng-Treitler Q. A model-agnostic approach for understanding heart failure risk factors. *BMC Res Notes* 2021; 14: 184.

72. Kheyfets V, Sweatt A, Gomberg-Maitland M, et al. Computational platform for doctor-artificial intelligence cooperation in pulmonary arterial hypertension prognostication: a pilot study. *ERJ Open Res* 2023; 9: 1–29.
73. Sun R, Wang X, Jiang H, et al. Prediction of 30-day mortality in heart failure patients with hypoxic hepatitis: development and external validation of an interpretable machine learning model. *Front Cardiovasc Med* 2022; 9: 1035675.
74. Shi H, Yang D, Tang K, et al. Explainable machine learning model for predicting the occurrence of postoperative malnutrition in children with congenital heart disease. *Clin Nutr* 2022; 41: 202–210.
75. Zhang X, Gavaldà R and Baixeries J. Interpretable prediction of mortality in liver transplant recipients based on machine learning. *Comput Biol Med* 2022; 151: 106188.
76. Xiang C, Wu Y, Jia M, et al. Machine learning-based prediction of disability risk in geriatric patients with hypertension for different time intervals. *Arch Gerontol Geriatr* 2023; 105: 104835.
77. Dong Z, Wang Q, Ke Y, et al. Prediction of 3-year risk of diabetic kidney disease using machine learning based on electronic medical records. *J Transl Med* 2022; 20: 143.
78. Dimitsaki S, Gavriilidis GI, Dimitriadis VK, et al. Benchmarking of machine learning classifiers on plasma proteomic for COVID-19 severity prediction through interpretable artificial intelligence. *Artif Intell Med* 2023; 137: 102490.
79. Ikemura K, Bellin E, Yagi Y, et al. Using automated machine learning to predict the mortality of patients with COVID-19: prediction model development study. *J Med Internet Res* 2021; 23: e23458.
80. Babaei Rikan S, Sorayaie Azar A, Ghafari A, et al. COVID-19 diagnosis from routine blood tests using artificial intelligence techniques. *Biomed Signal Process Control* 2022; 72: 103263.
81. Davazdahemami B, Zolbanin H and Delen D. An explanatory machine learning framework for studying pandemics: the case of COVID-19 emergency department readmissions. *Decis Support Syst* 2022; 161: 113730.
82. Pezoulas V, Kourou K, Papaloukas C, et al. A multimodal approach for the risk prediction of intensive care and mortality in patients with COVID-19. *Diagnostics (Basel, Switzerland)* 2021; 12. DOI: 10.3390/diagnostics12010056.
83. Pfaff E, Girvin A, Bennett T, et al. Identifying who has long COVID in the USA: a machine learning approach using N3C data. *Lancet Digit Heal* 2022; 4: e532–e541.
84. Alves M, Castro G, Oliveira B, et al. Explaining machine learning based diagnosis of COVID-19 from routine blood tests with decision trees and criteria graphs. *Comput Biol Med* 2021; 132: 104335.
85. Cavallaro M, Moiz H, Keeling M, et al. Contrasting factors associated with COVID-19-related ICU admission and death outcomes in hospitalised patients by means of Shapley values. *PLoS Comput Biol* 2021; 17: e1009121.
86. Davazdahemami B, Zolbanin H and Delen D. An explanatory analytics framework for early detection of chronic risk factors in pandemics. *Healthc Anal* 2022; 2: 100020.
87. Tang G, Luo Y, Lu F, et al. Prediction of sepsis in COVID-19 using laboratory indicators. *Front Cell Infect Microbiol* 2020; 10: 586054.
88. Zhang J, Sami S and Meiser-Stedman R. Acute stress and PTSD among trauma-exposed children and adolescents: computational prediction and interpretation. *J Anxiety Disord* 2022; 92: 102642.
89. Liu H, Dai A, Zhou Z, et al. An optimization for postpartum depression risk assessment and preventive intervention strategy based machine learning approaches. *J Affect Disord* 2023; 328: 163–174.
90. Schultebraucks K, Shalev A, Michopoulos V, et al. A generalized predictive algorithm of posttraumatic stress development following emergency department admission using biological markers routinely collected from electronic medical records. *Biol Psychiatry* 2020; 87: S101–S102.
91. Liu S, Schlesinger J, McCoy A, et al. New onset delirium prediction using machine learning and long short-term memory (LSTM) in electronic health record. *J Am Med Inform Assoc* 2022; 30: 120–131.
92. Banerjee S, Lio P, Jones P, et al. A class-contrastive human-interpretable machine learning approach to predict mortality in severe mental illness. *NPJ Schizophr* 2021; 7: 60.
93. Zhu T, Jiang J, Hu Y, et al. Individualized prediction of psychiatric readmissions for patients with major depressive disorder: a 10-year retrospective cohort study. *Transl Psychiatry* 2022; 12: 170.
94. Chmiel F, Burns D, Azor M, et al. Using explainable machine learning to identify patients at risk of reattendance at discharge from emergency departments. *Sci Rep* 2021; 11: 21513.
95. Twick I, Zahavi G, Benvenisti H, et al. Towards interpretable, medically grounded, EMR-based risk prediction models. *Sci Rep* 2022; 12: 9990.
96. Berezo M, Budman J, Deutscher D, et al. Predicting chronic wound healing time using machine learning. *Adv Wound Care* 2022; 11: 281–296.
97. Grazal C, Anderson A, Booth G, et al. A machine-learning algorithm to predict the likelihood of prolonged opioid use following arthroscopic hip surgery. *Arthroscopy* 2022; 38: 839–847.e2.
98. Valliani A, Kim N, Martini M, et al. Robust prediction of non-home discharge after thoracolumbar spine surgery with ensemble machine learning and validation on a nationwide cohort. *World Neurosurg* 2022; 165: e83–e91.
99. Zhang H, Wang Z, Tang Y, et al. Prediction of acute kidney injury after cardiac surgery: model development using a Chinese electronic health record dataset. *J Transl Med* 2022; 20: 166.
100. Wu Z, Li Y, Lei J, et al. Developing and optimizing a machine learning predictive model for post-thrombotic syndrome in a longitudinal cohort of patients with proximal deep venous thrombosis. *J Vasc Surgery Venous Lymphat Disord* 2023; 11: 555–564.e5.
101. Yoon J, Kim Y, Lee E, et al. Systemic factors associated with 10-year glaucoma progression in South Korean population: a single center study based on electronic medical records. *Sci Rep* 2023; 13: 530.
102. An C, Yang H, Yu X, et al. A machine learning model based on health records for predicting recurrence after microwave ablation of hepatocellular carcinoma. *J Hepatocell Carcinoma* 2022; 9: 671–684.

103. Jakobsen R, Nielsen T, Leutscher P, et al. Clinical explainable machine learning models for early identification of patients at risk of hospital-acquired urinary tract infection. *J Hosp Infect* 2023: 23. doi:10.1016/j.jhin.2023.03.017
104. Liu M, Guo C and Guo S. An explainable knowledge distillation method with XGBoost for ICU mortality prediction. *Comput Biol Med* 2023; 152: 106466.
105. Dong S, Khattak A, Ullah I, et al. Predicting and analyzing road traffic injury severity using boosting-based ensemble learning models with SHAPley additive exPlanations. *Int J Environ Res Public Health* 2022; 19. DOI: 10.3390/ijerph19052925
106. El-Sappagh S, Alonso J, Islam S, et al. A multilayer multimodal detection and prediction model based on explainable artificial intelligence for Alzheimer's disease. *Sci Rep* 2021; 11: 2660.
107. Kavalci E and Hartshorn A. Improving clinical trial design using interpretable machine learning based prediction of early trial termination. *Sci Rep* 2023; 13: 121.
108. Chun M, Park C, Kim J, et al. Prediction of conversion to dementia using interpretable machine learning in patients with amnestic mild cognitive impairment. *Front Aging Neurosci* 2022; 14: 898940.
109. Qiu W, Chen H, Dincer AB, et al. Interpretable machine learning prediction of all-cause mortality. *Commun Med* 2022; 2: 125.
110. Xu J, Chen X and Zheng X. Acinetobacter baumannii complex-caused bloodstream infection in ICU during a 12-year period: predicting fulminant sepsis by interpretable machine learning. *Front Microbiol* 2022; 13: 1037735.
111. Lundberg S, Nair B, Vavilala M, et al. Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. *Nat Biomed Eng* 2018; 2: 749–760.
112. Mohanty S, Lekan D, McCoy T, et al. Machine learning for predicting readmission risk among the frail: explainable AI for healthcare. *Patterns (New York, NY)* 2022; 3: 100395.
113. Shim S, Yu J, Jekal S, et al. Development and validation of interpretable machine learning models for inpatient fall events and electronic medical record integration. *Clin Exp Emerg Med* 2022; 9: 345–353.
114. Meiseles A, Paley D, Ziv M, et al. Explainable machine learning for chronic lymphocytic leukemia treatment prediction using only inexpensive tests. *Comput Biol Med* 2022; 145: 105490.
115. Javed A, Khan H, Alomari M, et al. Toward explainable AI-empowered cognitive health assessment. *Front Public Heal* 2023; 11: 1024195.
116. Vyas A, Aisopos F, Vidal M-E, et al. Identifying the presence and severity of dementia by applying interpretable machine learning techniques on structured clinical records. *BMC Med Inform Decis Mak* 2022; 22: 271.
117. Thorsen-Meyer H, Nielsen A, Nielsen A, et al. Dynamic and explainable machine learning prediction of mortality in patients in the intensive care unit: a retrospective study of high-frequency data in electronic patient records. *Lancet Digit Heal* 2020; 2: e179–e191.
118. Ke X, Zhang F, Huang G, et al. Interpretable machine learning to optimize early in-hospital mortality prediction for elderly patients with sepsis: a discovery study. *Comput Math Methods Med* 2022; 2022: 4820464.
119. Ning Y, Li S, Ong M, et al. A novel interpretable machine learning system to generate clinical risk scores: an application for predicting early mortality or unplanned readmission in a retrospective cohort study. *PLOS Digit Heal* 2022; 1: e0000062.
120. Yang B, Xu S, Wang D, et al. ACEI/ARB medication during ICU stay decrease all-cause in-hospital mortality in critically ill patients with hypertension: a retrospective cohort study based on machine learning. *Front Cardiovasc Med* 2021; 8: 787740.
121. Choi T, Chang M, Heo S, et al. Explainable machine learning model to predict refeeding hypophosphatemia. *Clin Nutr ESPEN* 2021; 45: 213–219.
122. Alsinglawi B, Alshari O, Alorjani M, et al. An explainable machine learning framework for lung cancer hospital length of stay prediction. *Sci Rep* 2022; 12: 607.
123. Zhang C, Zhang Y, Yang Y, et al. Machine learning models for predicting one-year survival in patients with metastatic gastric cancer who experienced upfront radical gastrectomy. *Front Mol Biosci* 2022; 9: 937242.
124. Lu S, Knafl M, Turin A, et al. Machine learning models using routinely collected clinical data offer robust and interpretable predictions of 90-day unplanned acute care use for cancer immunotherapy patients. *JCO Clin Cancer Informatics* 2023; 7: e2200123.
125. Caicedo-Torres W and Gutierrez J. ISeeu: visually interpretable deep learning for mortality prediction inside the ICU. *J Biomed Inform* 2019; 98: 103269.
126. Singh D, Nagaraj S, Mashouri P, et al. Assessment of machine learning-based medical directives to expedite care in pediatric emergency medicine. *JAMA Netw Open* 2022; 5: e222599.
127. Jacobson D, Cadieux B, Higano C, et al. Risk factors associated with skeletal-related events following discontinuation of denosumab treatment among patients with bone metastases from solid tumors: a real-world machine learning approach. *J Bone Oncol* 2022; 34: 100423.
128. Levy MD, Loy L and Zatz LY. Policy approach to nutrition and physical activity education in health care professional training. *Am J Clin Nutr* 2014; 99: 1194S–1201S.
129. Xie F, Chakraborty B, Ong M, et al. Autoscore: a machine learning-based automatic clinical score generator and its application to mortality prediction using electronic health records. *JMIR Med Informatics* 2020; 8: e21798.
130. Slack D, Hilgard S, Jia E, et al. *Fooling LIME and SHAP*. In: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, New York, NY, USA: ACM, pp.180–186.