Federated Learning-Based Model for Predicting Mortality: Systematic Review and Meta-Analysis

Nurfaidah Tahir, Chau-Ren Jung, Shin-Da Lee, Nur Azizah, Nur Azizah, Tsai-Chung Li

Submitted to: Journal of Medical Internet Research
on: August 23, 2024

Table of Contents

Original Manuscript **5**
Supplementary Files **22**
Multimedia Appendixes **23**
Multimedia Appendix 1 ..... **23**
Multimedia Appendix 2 ..... **23**
Multimedia Appendix 3 ..... **23**
Multimedia Appendix 4 ..... **23**
Multimedia Appendix 5 ..... **23**

Federated Learning-Based Model for Predicting Mortality: Systematic Review and Meta-Analysis

Nurfaidah Tahir[1, 2, 3]; Chau-Ren Jung[1]; Shin-Da Lee[4]; Nur Azizah[1]; Nur Azizah[1]; Tsai-Chung Li[1]

[1]Department of Public Health, China Medical University Taichuang TW
[2]PhD Program in Department of Public Health, China Medical University Taichuang TW
[3]Department of Industrial Engineering, Hasanuddin University Makassar ID
[4]Department of Physical Therapy, China Medical University Taichuang TW

Abstract

Background: The quality of a machine learning model considerably relies on the size of the dataset, the development and widespread application of this method have often been hindered by confidentiality issues, particularly regarding data privacy. Predicting mortality is essential in clinical environments. When a patient is admitted, estimating their likelihood of mortality by the end of their intensive care unit (ICU) stay or within a designated time frame is a way to assess the severity of their condition. This information is crucial in managing treatment planning and resource allocation. However, individual hospitals typically have a limited amount of local data available to create a reliable model. The rise of federated learning as a novel privacy-preserving technology offers the potential for collaboratively creating models in a decentralized manner, eliminating the need to consolidate all datasets in a single location. Nonetheless, there is a scarce of clear and comprehensive evidence that compares the performance of federated learning with that of traditional centralized machine learning approaches, particularly considering healthcare implementation.

Objective: This study aims to review the comparison of performances between federated learning (FL)-based and centralized machine learning (CML) models for mortality prediction in clinical settings.

Methods: The electronic database search was conducted for English articles that developed federated-based learning model to predict mortality. Screening, data extraction, and risk of bias assessments were carried out by at least two independent reviewers. Meta-analyses of pooled area under the receiver operating curve (AUROC/AUC) values were examined for FL, CML, and LML. The risk of bias was assessed using critical appraisal and data extraction for systematic reviews of prediction modeling studies (CHARMS) and prediction model risk of bias assessment tool (PROBAST) guidelines

Results: In total, 9 articles that were heterogeneous in framework design, scenario, and clinical context were included (n = 5 [55.6%] were observed in specific case; n = 3 [33.0%] were in ICU settings; and n = 2 [22.0%] in emergency department, urgent, or trauma center). Cohort datasets were utilized by all included studies. These studies universally indicated that performance of FL model outperforms LML model and closest to the CML model. The pooled AUC for FL and, CML (or LML) performances were 0.81 (95 % CI 0.76–0.85, I2 78.36 %) and 0.82 (95 % CI 0.77–0.86, I2 72.33 %), respectively. All included studies had either a low, high, or unclear risk of bias.

Conclusions: This systematic review and meta-analysis demonstrate that federated learning models outperform local machine learning approaches and are comparable to centralized models. However, efficiency may be compromised due to complexity, privacy preservation, and high computation and communication costs. Clinical Trial: PROSPERO International Prospective Register of Systematic Reviews CRD42024539245; https://www.crd.york.ac.uk/prospero/display_record.php?RecordID=539245

(JMIR Preprints 23/08/2024:65708)
DOI: https://doi.org/10.2196/preprints.65708

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Please make my preprint PDF available to anyone at any time (recommended).**
Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users. Only make the preprint title and abstract visible.

No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**

Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v

Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http

Original Manuscript

Review

Nurfaidah Tahir[1,2,3], MSc; Chau-Ren Jung[1], PhD; Shin-Da Lee[1,4], PhD; Nur Azizah[1], MD; Wen-Chao Ho[1], PhD; Tsai-Chung Li[1], PhD
Department of Public Health, China Medical University, Taiwan
PhD Program in Department of Public Health, China Medical University, Taiwan
Department of Industrial Engineering, Hasanuddin University, Makassar, Indonesia
Department of Healthcare, China Medical University, Taiwan

Corresponding Author
Wen-Chao Ho, PhD
Department of Public Health Professor
China Medical University
Jingmao Road, Beitun District, Taichung City
406040
Taiwan
Email: wcho@mail.cmu.edu.tw

Federated Learning-Based Model for Predicting Mortality: Systematic Review and Meta-Analysis

Abstract

Background: The quality of a machine learning model considerably relies on the size of the dataset, the development and widespread application of this method have often been hindered by confidentiality issues, particularly regarding data privacy. Predicting mortality is essential in clinical environments. When a patient is admitted, estimating their likelihood of mortality by the end of their intensive care unit (ICU) stay or within a designated time frame is a way to assess the severity of their condition. This information is crucial in managing treatment planning and resource allocation. However, individual hospitals typically have a limited amount of local data available to create a reliable model. The rise of federated learning as a novel privacy-preserving technology offers the potential for collaboratively creating models in a decentralized manner, eliminating the need to consolidate all datasets in a single location. Nonetheless, there is a scarce of clear and comprehensive evidence that compares the performance of federated learning with that of traditional centralized machine learning approaches, particularly considering healthcare implementation.
Objective: This study aims to review the comparison of performances between federated learning (FL)-based and centralized machine learning (CML) models for mortality prediction in clinical settings.
Methods: The electronic database search was conducted for English articles that developed federated-based learning model to predict mortality. Screening, data extraction, and risk of bias assessments were carried out by at least two independent reviewers. Meta-analyses of pooled area under the receiver operating curve (AUROC/AUC) values were examined for FL, CML, and LML. The risk of bias was assessed using critical appraisal and data extraction for systematic reviews of prediction modeling studies (CHARMS) and prediction model risk of bias assessment tool (PROBAST) guidelines
Results: In total, 9 articles that were heterogeneous in framework design, scenario, and clinical context were included (n = 5 [55.6%] were observed in specific case; n = 3 [33.0%] were in ICU settings; and n = 2 [22.0%] in emergency department, urgent, or trauma center). Cohort datasets were utilized by all included studies. These studies universally indicated that performance of FL model outperforms LML model and closest to the CML model. The pooled AUC for FL and, CML (or

LML) performances were 0.81 (95 % CI 0.76–0.85, I2 78.36 %) and 0.82 (95 % CI 0.77–0.86, I2 72.33 %), respectively. All included studies had either a low, high, or unclear risk of bias.
Conclusions: This systematic review and meta-analysis demonstrate that federated learning models outperform local machine learning approaches and are comparable to centralized models. However, efficiency may be compromised due to complexity, privacy preservation, and high computation and communication costs. **Trial Registration:** PROSPERO International Prospective Register of Systematic Reviews CRD42024539245; https://www.crd.york.ac.uk/prospero/display_record.php?RecordID=539245

Keywords: federated learning; centralized machine learning; mortality prediction.

Introduction

The emergence of machine learning (ML) as a subset of artificial intelligence (AI) contributed to the novel of computational thinking. ML empowers computers to “learn” from training data and augment their knowledge without the need for explicit programming. The algorithm of ML able to find patterns from data and use that knowledge to generate their own prediction. In concise, machine learning models and algorithm acquire knowledge through experience. Typically, a set of instructions have been given to developed computer program by engineers to turn incoming data into its intended output. In contrast, ML program is designed to learn with minimum or no human intervention and to broad the knowledge over time. Researches in various domain are drawn to ML due to its vast potential in classification and regression problems, and ability to apply both supervised and unsupervised learning approaches. Previous studies demonstrated the range of ML applications can be discovered in the field such as: User behavior analytics and context-aware smartphone applications [1, 2]; Image, speech and pattern recognition [1, 2]; E-commerce and product recommendations [1, 2]; Traffic prediction and transportation [1, 3]; Healthcare services [4, 5]; Cybersecurity and threat intelligence [6]; Internet of Things (IoT) and smart cities [3]; Sustainable agriculture [7]; Industrial applications [8]; and Natural language processing and sentiment analysis [9].

Machine Learning and Its Challenge

Accuracy results obtained from ML classification or regression tasks encourage the utilization of these approaches in areas of daily life. The promise of ML models has been demonstrated by the precision they offer and the potential to incorporate them across different fields. However, despite its successful implementation, ML still suffers from several challenges. The primary steps in the machine learning pipeline consist of data collection and preprocessing, feature engineering, model training, model evaluation, and model deployment. The importance of data is explicitly illustrated in the structure of the ML workflow. In short, the performance of ML models highly depends on the availability of data. While the technical structure, data cleanliness, feature selection, and other factors are important to achieving highly accurate models, it is widely recognized that having more data for training is crucial to improving model accuracy [10, 11]. However, in practice, gathering data for ML models is one of the major challenges when it comes to privacy and confidentiality.

Since society, organizations, and governments are strengthening the security and privacy protections for data, several laws and regulations have been enacted, for instance China’s Cyber Security Law of the People’s Republic of China [12], the Personal Data Protection Act (PDPA) in Singapore [13], the European Union’s General Data Protection Regulation (GDPR), and many other principles legislated around the globe. Although these restrictions aid in the protection of private information, they present additional difficulties for the ML community to gather the data for model training, which in

turn makes it more difficult to enhance the performance accuracy and personalization of those models. Therefore, data privacy and confidentiality are not isolated problems; rather, they trigger other issues with ML, including data availability, performance, customization, and therefore acceptability and trust.

Machine Learning for Mortality Prediction

Predicting mortality is essential in clinical environments. When a patient is admitted, estimating their likelihood of mortality by the end of their intensive care unit (ICU) stay or within a designated time frame is a way to assess the severity of their condition. This information is crucial in managing treatment planning and resource allocation [14]. However, individual hospitals typically have a limited amount of local data available to create a reliable model. Typically, a healthcare institution would implement a domain transfer of an existing in-hospital mortality prognostic model that has been constructed utilizing publicly available datasets [15]. The sharing of additional datasets from various healthcare facilities can significantly enhance both the performance and generalizability of these models [16]. This underscores the critical role of data sharing in the advancement of high-performing predictive models within clinical environments.

However, within the healthcare sector, it is quite prevalent for hospitals to isolate their datasets—often justifying this practice with legitimate privacy concerns—while undertaking internal model development [17]. Despite the hospitals' belief in the benefits of data sharing, conducting analysis in a centralized fashion, which necessitates the consolidation of datasets from all participating hospitals or centers, heightens the risks associated with data privacy and security, as sensitive information is now disseminated to outside entities. Furthermore, the transfer of datasets to a centralized repository, whether through physical means or via network channels, creates an additional vulnerability for potential data breaches [18, 19]. In addition to the privacy and security challenges, the administrative burden of orchestrating data sharing is also significant, as each participant typically adheres to its own regulations concerning data utilization and ownership [19]. Consequently, a methodology that facilitates collaborative modeling in a decentralized framework, eliminating the requirement to aggregate all datasets in a singular location, would significantly enhance the feasibility of multi-center studies.

Proposing Federated Learning (FL)

The new concept in the ML domain known as federated machine learning, or federated learning (FL), was introduced by Google in 2016 [20]. The architecture of FL proposes to eliminate the data exchange between the participants. As a type of collaboratively distributed or decentralized ML privacy-preserving technology, FL eliminates the need to transfer data from the nodes to a central server. The principle of FL, or client-based architecture, enables multiple institutions to collaborate, wherein the baseline model will be hosted by a coordinating node while computational nodes download the model and train it on local datasets. After completing the training process, client parameter weights will be sent to the central server for aggregation. The client's own dataset remains unshared and cannot be accessed or manipulated by third parties. Through the continuous repetition of this process, the effectiveness of the model can be enhanced and incorporate newly available data, facilitating the advancement of continuously updated real-world evidence-based knowledge [15, 20].

Federated learning has demonstrated itself to be a great solution to address privacy issues, enabling large data sets to be trained using ML-based models and enhancing their precision and effectiveness. In addition, FL attempts to formulate models from various data sets and merge the knowledge into a global trained model, which increases the efficiency of the models. Moreover, this strategy is effective for investigating medical conditions [21], particularly those with scarce prevalence or

minimal data, to prevent inadequate care resulting from misrepresenting or underrepresenting certain patient groups [22].

Federated Learning Borderline

The primary goal of both FL and classical ML is to optimize the learning goal. However, they differ in the architecture and design of their models. Classical ML, known as centralized ML (CML), is the concept where data characterized by the same features is collected from different users to be localized on a central server, where it is then processed and analyzed. This concept is similar to the local ML approach; the differences are only in terms of the server. While CML incorporated data from all users, LML used their local data to be processed and analyzed at their sites. In this context, these concepts can be compared in terms of: motivation; data identity; centralization; data access; communication; and data transfer [15]. that is presented in Table 1.

Table 1. Borderline between FL, CML, and LML

	Centralized Machine Learning	Local Machine Learning	Federated Learning
Privacy	Not Considered	Not Considered	Main objective
Data Identity	Independent and Identically Distributed (IID)	Independent and Identically Distributed (IID)	Non-IID supported
Centralization	Centralized to one server	Local site	Only Aggregation
Access to Data	Main server has full access	Local server	No access
Communication and Data Transmission	All data transmitted	All data transmitted	Only parameters

The promise of FL in healthcare is depicted by its ability to enable model training using distributed and decentralized health data. Facilitating the creation of a high-accuracy and more personalized model by utilizing a myriad of datasets while maintaining the privacy of the patient is the main objective of the FL framework. Moreover, FL has enhanced its ability to learn by analyzing data that are distributed among multiple sites and cannot be merged into one dataset, or when data are fragmented across multiple clinical systems [15]. In summary, healthcare quality can be improved by leveraging FL to shift towards a more data-driven and personalized approach.

Outline and Main Contribution of This Article

In this article, FL and its use in mortality prediction have been studied. The primary objective of this systematic review and meta-analysis is to compare the performance of mortality prediction models developed or validated using the FL approach to those developed using classical ML methods (centralized and local). The receiver operating characteristic (AUROC) or area under curve (AUC) value will be interpreted as an evaluation metric of the developed models. Comparisons will only be made between models utilizing AUROC or AUC for the evaluation metrics. Our secondary objective is to compare the resources (e.g., training and prediction time) among these health data models using different architectures. In this context, this article attempts to answer the following questions: “In different clinical contexts, how is the feasibility and capability of a federated learning approach for predicting mortality when compared to centralized or local machine learning?”.

Methods

Design

This systematic review and meta-analysis is reported in accordance with the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidance [23] (Multimedia Appendix 1).

Eligibility Criteria

Articles published after 2016, were eligible for inclusion if they investigated research evidence targeted at clinical context. Articles were included if they quantitatively compared FL versus CML or LML model in predicting mortality (before-after federated within-group comparisons) in the terms of AUROC/AUC. Eligible study designs included experimental study that compare the performance of FL and CML or LML. Eligible outcome of interest is mortality prediction.

Articles were not eligible for inclusion if they did not compare FL to CML or LML or only compared the effectiveness of FL performance in different FL-baseline model. Excluded study designs included protocols, reviews, studies using only qualitative methods, opinion pieces, and conference abstracts with no linked full-text article. Articles were also excluded if they are evaluating the model performance with another evaluation metrics instead of AUROC/AUC and if they were not available in English.

Information Sources

An actual search was conducted in four multi-disciplinary databases (IEEE Xplore, PubMed, Science Direct, Web of Science) with the assistance of EndNote 20 software. The date of the last search was June 23, 2024. Manual searches of the reference lists, citations, and related articles of the included studies were undertaken to identify additional studies missed from the original electronic searches.

Search Strategy

The controlled free-text terms were used through the Boolean operators (Multimedia Appendix 2). All original studies that developing mortality prediction model were included if they met the predefined inclusion criteria (PICO):

- Population: Patients in different clinical setting (e.g. ICU, emergency department [ED]), and Trauma Center, or specific disease admission).
- Intervention: Federated learning model.
- Comparator: Centralized or local machine learning model.
- Outcomes: Mortality prediction.

Selection Process

Records from the electronic and citation searches were exported to EndNote Online (Clarivate) for deduplication and title, abstract, and full-text screening. Title and abstract screening were carried out by two independent reviewers (NT and NA). Full-text screening was carried out by at least two of the six independent reviewers (NT, CRJ, SDL, NA, WCH, and TCL). In cases of disagreement, conflict was solved by discussion with each other until 100% agreement was achieved.

Data Collection

Data from the included studies were extracted independently by at least two of the six reviewers (see the Selection Process section) using a data extraction form developed a priori. The accuracy of data extraction was confirmed by comparison between extraction forms, returning to the original article to resolve any disparity.

Data Items

The variables collected were study characteristics including the data source, number and description of subjects; study design; comparisons; and outcome. For outcome of interest, AUROC/AUC, variance, and sample sizes were extracted for each comparison. When these data were missing, they were calculated from other reported statistics using recommended methods [24], where possible. For studies that reported multiple outcome measures, only outcome of interest was collected (mortality prediction).

Risk of Bias Assessment

Two (NT and NA) of six reviewers independently assessed the risk of bias using the Prediction Model Risk of Bias Assessment Tool (PROBAST) [25]. In case of disagreement, conflict was solved by discussion with other researchers. The PROBAST includes 20 signaling questions across four key domains (participants, predictors, outcome, analysis), while each domain is assessed for a low, high, or unclear risk of bias. Subtype 1 of the Critical Appraisal and Data Extraction for Systematic Reviews of Prediction Modeling Studies (CHARMS), prediction model development without external validation, also examined in conjunction with the PROBAST tool [26] by at least 2 of 6 reviewers.

Data Synthesis

The included studies were summarized narratively in text, tables, and figures. Discrimination (model's ability to differentiate between mortality and survived) was extracted to estimate prediction models' ability to distinguish survived patients and death event (range from 0.5-no discriminative ability to 1-perfect discriminative ability) [24]. Due to lack of calibration plots, summarization of calibration (the agreement between the frequency of observed events with the predicted probabilities) was not assessed. Prognostic prediction models with effect sizes (AUROCs) for the same outcome were combined and utilized in a meta-analysis using R package. Standard errors were estimated based on normal distribution assumption. Because the included studies typically differ in design, and execution (Multimedia Appendix 3), variation between their results are unlikely to occur by chance only. For this reason, the meta-analysis should usually allow for the presence of heterogeneity and aim to produce a summary result (with its 95% confidence interval) that quantifies the average performance across studies. This can be achieved by implementing a random (rather than a fixed) effects meta-analysis model [23, 24]. In addition, the Higgins I^2 test was used to evaluate the heterogeneity between the included studies ($I^2 \leq 25\%$ for low, $I^2 < 50\%$ for moderate, and $I^2 \geq 50$ % for high) (Multimedia Appendix 4).

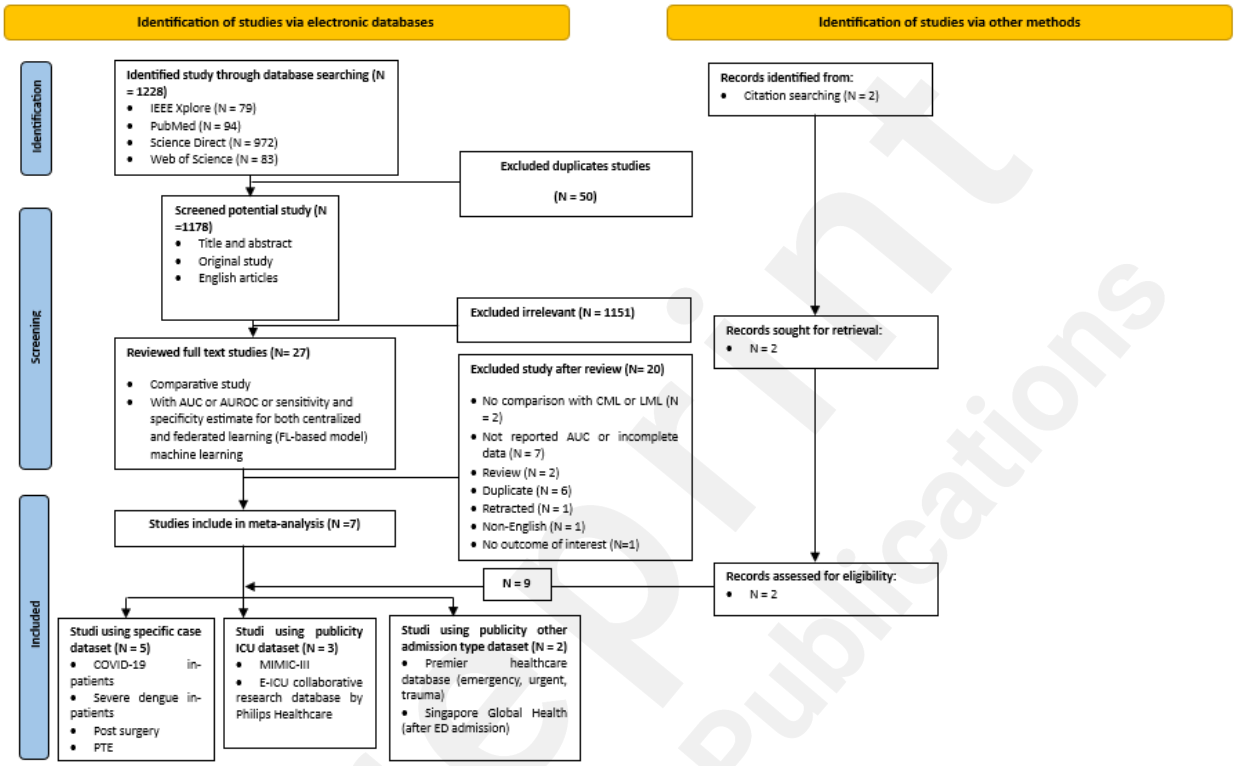
Certainty Assessment

For each outcome, the performance of FL and CML or LML model was evaluated using C statistic. When measures of uncertainty were not reported, we approximated the standard error of the C statistic by using the appropriate and suggested measurement (Multimedia Appendix 4).

Results

In total, 1228 records were identified, 29 full-text reports were screened, and 9 articles were included (Figure 1).

Figure 1. Preferred Reporting Items for Systematic Reviews and Meta-Analyses(PRISMA) flow diagram illustrating the process of study selection for a systematic review on the federated learning-based model for predicting mortality, detailing a total of 1228 records identified (1201 excluded), 29 full-text reports screened (20 excluded), and 9 articles included.



Included Studies

Study selection was performed in three stages: At the first stage (N = 1228), the studies were screened for duplicate hits with the assistance of EndNote 20 software. At the second stage, potentially relevant studies were assessed by comparing the titles and abstracts (n = 1178) against the predetermined inclusion criteria. At the third stage, the studies (n = 27) that appeared to meet the inclusion criteria and 2 articles that sought from the citation were obtained for detailed assessment.

Among the 29 studies identified and assessed for eligibility, 20 were excluded due to no comparison with CML or LML or only included the other federated learning model as the comparator group, using other evaluation metrics instead of AUROC/AUC value, review articles, and retracted articles. Majority (89%) of included studies were retrospective cohort studies using institutional [27, 28, 29, 30, 31, 32, 33, 35] data sources. The follow-up period ranged from 2001-2021, and the longest followed-up period was 11 years. Prediction models developed in each study using internal validation. Across the nine studies, a total two machine learning based models (neural network and logistic regression), were developed and validated. For study descriptions, refer to Multimedia Appendix 3.

Risk of Bias

Most of the studies had retrospective cohort design and only one prospectively analyzed electronic record data. The optimal design for constructing a predictive model for a future event is to use a

cohort in which data are obtained prospectively [36]. This condition was satisfied by Vaid model. Although most of the cohort data utilized by included studies were obtained using a random sample design, one study did not provide any information related to missing value and data quality check [28]. Thus, the sample may not be representative of the target population and this study rated as unclear risk of bias in its participant domain.

Three studies defined the outcome as the death of patient in ICU [27, 28, 32], four studies defined mortality as the death of patient due to specific condition (COVID-19, post visceral surgery, and pulmonary thromboembolism [PTE]), and two studies used the number of death patient in ED as the outcome to be predicted [29, 30]. Some of the studies determined the time of outcome occurrence with a widely accepted time interval (i.e. 28h, 48h, 7 days, 30 days) [37]. However, the information regarding the time of outcome occurrence cannot be found in three studies [28, 31, 33]. In the final assessment, two studies [31, 33] rated as high risk of bias and unclear concern for applicability because they did not have either the information regarding the time interval of predictor assessment or outcome occurrence.

All the studies gave the number and type of predictors. Furthermore, all the predictors were measured at certain baseline and were therefore measured at the same time in all patients, thus minimizing information bias. Concerning the form of measurement, we note that since mortality prediction used data from four different cohorts, there are differences in variables used to be predictors and form of measurement. Most of studies use demographic information, vital signs, and lab results as the predictors, while only one study used drug features [29] and CT images [33]. All continuous variables were measured and handled appropriately, except one study conducted by [30], wherein the numerical features were transformed into categorical. However, the creation of categorical variables allows for the modeling of nonlinear effects [25], which has been widely applied in the development of clinical scoring systems. Regarding the blinding in the measurement of the predictors, only one study [33] informed the masking of predictors measurement. However, the lack of information on blinding or masking predictor evaluation did not immediately offer the prospect of bias because all predictors were objective.

The total number of subjects wherein the models were developed was between 3,055 and 1,222,554 subjects. These figures were between 176 and 38,666 for the number of cases of mortality. Prediction models constructed with machine-learning techniques frequently necessitate considerably elevated event per variable (EPV) ratios (often exceeding 200) to mitigate the risk of overfitting. While certain investigations have reported lower EPVs, the investigators assess the degree of misfitting present in the established prediction model (for instance, by employing internal validation methodologies). Through this internal validation process, optimism-adjusted performance estimates of the model can be derived, and model parameters can be recalibrated (that is, shrink regression coefficients) to attenuate this bias [25].

One study [30] explicitly mentioned the used of complete-case analysis to address missing value, without indicating the number of missing subjects in each variable, while three studies [28, 29, 33] gave no any information of their missing value. Participants with incomplete data are likely excluded from analyses (termed “available-case” or “complete-case” analysis) since statistical software typically disregards individuals with any absent values in the analyzed dataset, unless prompted to handle otherwise [25]. Employing complete-case analysis may potentially introduce bias into the findings; thus, it is advisable to utilize multiple imputation techniques for addressing missing data, unless the missing data have a missing completely at random pattern [36]. Consequently, four included studies have a high potential of bias in the analysis domain.

All the studies used either neural network or logistic regression as the baseline model. Regarding the predictors to be considered in the multivariate model, in all the articles the authors selected the significant predictors, either unadjusted or adjusted for age and sex. After selecting the candidate variables to be included in the multivariate model, four studies used machine learning method for selecting candidate predictors: dropout method [27]; model averaging mechanism [28]; parsimony plot [30]; multivariate time series (MTS) [32]. Two other studies used regression technique (LASSO regression and ridge regression) [33, 34], and three studies without information [29, 31, 35]. About a half of the studies employed shrinkage techniques [27, 28, 31, 33, 34]. Finally, none studies reported the selecting predictors according to the univariate associations.

All the models assessed the discrimination but none of study assessed the calibration. Regarding discrimination, the authors mainly used the AUC or the C-statistic. Most of the studies developing a prediction model by utilizing a full cohort approach (without sampling) and therefore did not involve follow-up, censoring, or competing events. The included studies combined the randomized split-sample technique and bootstrapping cross-validation technique or optimizer method to evaluated the model. However, clinical utility and external validation to ensure the generalization of the results were not taking into consideration by any of study. The detail information regarding the result of risk of bias assessment by CHARMS and PROBAST guideline is provide in tabulation manner and can be seen in Multimedia Appendix 5.

Federated Learning and Mortality Prediction: State-of-the-Art

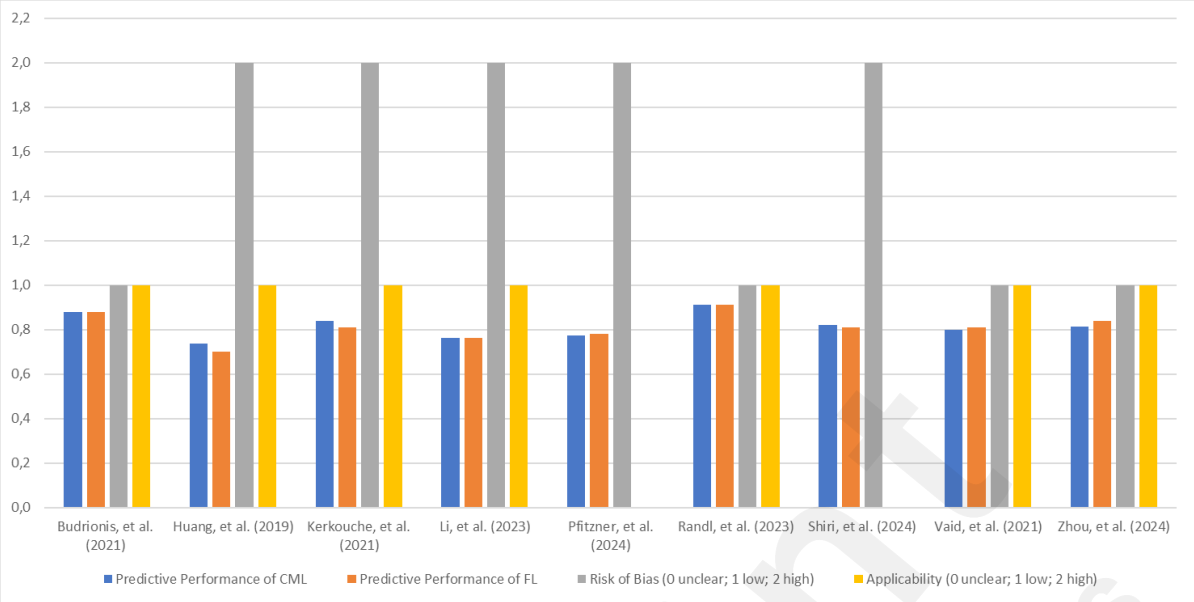
A study developed Feedforward Neural Network (FNN) combined with Recurrent Neural Network (RNN) as the baselined machine learning model to predict mortality among ICU patients [27]. This study aims to benchmark the FL models performance in three different configurations setting: data, nodes, and distribution, then compare them with CML model. The best approach is achieved by configuring the data setting (fixed on nodes and distribution form, while increasing the data). The result demonstrated that performance of FL was almost similar with CML in the terms of F1-Score and AUROC (F1-Score: CML= 0.37 and FL=0.30; ROC AUC: CML= 0.88 and FL =0.88). However, FL model training and inference took approximately 9 and 40 times longer, respectively, than the equivalent tasks executed in centralized settings. Extracting the same public dataset with previous study, MIMIC-III, [32] focused on the utilization of less resource-intensive Gated Recurrent Unit (GRU) as their basic deep learning RNN model to predict the risk of ICU mortality in an early stage. They evaluated the FL performance in two different schemes (using early stop with minimum loss and early stop with maximum F1-Score) and different amounts of clients (2, 4, and 8). After comparing the predictive performance of FL, CML, and LML in the terms of AUPRC, F1-score, and AUROC, the obtaining results show that FL model performs equally well as the centralized approach and was substantially better than the local approach. Another scheme constructed by [28], which introduced a community-based federated machine learning (CBFL) algorithm and evaluated it on non-IID ICU electronic health records (EHRs) data of medications extracted from e-ICU Collaborative Research Database. This study clustered the distributed data into clinically meaningful communities that captured similar diagnoses and geographical locations, and learnt one model for each community. Evaluation results show that CBFL outperformed the baseline federated machine learning (FL) and was not all that different from CML in terms of ROC AUC (CML = 0.7368, FL = 0.6895, CBFL10 = 0.6989).

Experimental evaluation of FL also implemented in emergency setting [29] by proposing FL sign differential privacy (FL-SIGN_DP) scheme to evaluate the performance of FL-based model on a realistic in-hospital mortality prediction scenario. In the FL-STANDARD scheme, each client sends

its updated model to the central server. In contrast, in FL-SIGN scheme, each client only sends the "sign" of coordinate value in its parameter update vector. The authors experimentally evaluate the performance of their solution for in-hospital mortality prediction using the Premier Healthcare database, collecting information from millions of patients over a period of 12 months from 415 hospitals in the United States. The accuracy performance results that FL-STANDARD outperform FL-SIGN and close to CML in the term of AUROC (CML = 0.84, FL-STANDARD =0.81, FL-SIGN = 0.77). In addition, the study [30] implemented scoring-based system (FedScore model) to facilitate cross-institutional collaborations in developing mortality prediction within 30 days after emergency department visit. The authors incorporated 10 simulated sites divided from a tertiary hospital in Singapore. This study found FedScore model achieved better performance with an AUC value of 0.7633 across all sites, and standard deviation (SD) of 0.0204, compare to CML (AUC = = 0.7631; SD = 0.0289) and the best LML (AUC = = 0.7627; SD = 0.0262).

Different from previous studies, by incorporating the retrospective patient's data from Department of Surgery, Campus Charit´ e Mitte, Campus Virchow Klinikum, Charit´ e–Universitatsmedizin Berlin, [31] investigated the relationship between FL, differential privacy (DP), and highly non-independent and identically distributed (non-IID) data for predicting patient mortality and revision surgery after visceral operations. The evaluation of local models without DP exhibits a similar utility as FL for the prognosis of revision surgery, but for mortality prediction, FL reached higher area under the precision-recall curves (AUPRCs). This author assumed that the smaller class imbalance for the former task allows local learning to perform better, but given a very small fraction of positive class samples, data owners can benefit a lot by collaborating with others. This study experimentally demonstrated that FL without DP perform better that LML and CML (FL = 0.779; LML = 0.756; and CML = 0.773), in terms of AUROC. One study conducted by [33] in 19 centers hospitals in Iran, evaluated the performance of deep privacy preserving federated learning (DPFL) and the differences between FL and CML performance in predicting COVID-19 outcomes using chest computed tomography (CT) images. In obtaining result, the FL with adaptive quantile clipping (GDP-AQuCl) that combined with DP outperformed centralized models while operating on large and heterogeneous multi-institutional datasets. In addition, the model was resistant to inference attacks, ensuring the privacy of shared data during the training process. The mean AUC of 0.82 and 0.81 with 95% confidence intervals of (95% CI:0.79–0.85) and (95% CI:0.77–0.84) were achieved by the centralized model and the DPFL model, respectively, and the DeLong test did not prove statistically significant differences between the two models (p-value = 0.98). Study by [34] is the only study used prospective HER dataset of 5 hospitals within the Mount Sinai Health System to predict mortality in hospitalized patients with COVID-19 within 7 days of follow-up. They developed Logistic regression with L1 regularization/least absolute shrinkage and selection operator (LASSO) and multilayer perceptron (MLP) models as the baseline ML model. This study identified that federated MLP outperformed centralized and local MLP model at all 5 hospitals, as determined by the mean of area under the receiver operating characteristic curve (MLP Model Mean AUROC from 5 Sites: CML = 0.80, LML = 0.76 FL = 0.81). Finally, with a logistic regression as the baseline model, [35] experimentally proofed that using real-world data, the FL model performed better than the centralized model in predicting PTE prognostic risk (AUC CML = 0.812; FL = 0.840). Moreover, the result also demonstrated that FL can improve the generalization performance of the model to some extent, reflecting successful decentralized optimization with diverse distributions of training data. FL showed that it effectively utilizes datasets from all clients, while each client only uses its own local dataset to train the model.

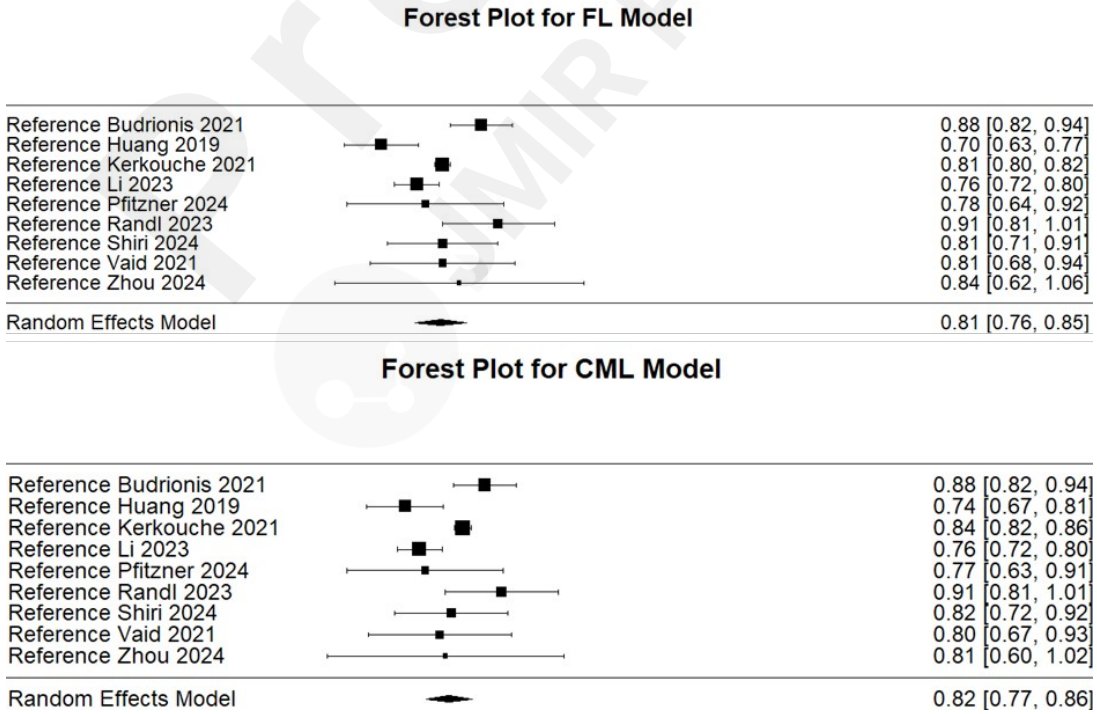
Figure 2. Summary of included studies by the definition of mortality prediction, predictive performance (AUROC), and the overall risk of bias and applicability.



Predictive Performance

Most studies used more than one evaluation metric related to discrimination (e.g., AUROC or AUC, sensitivity or recall, specificity, precision, accuracy, AUPRC, and F1-Score). Calibration performance was not evaluated due to lack of information. The pooled AUC for FL performance and CML or LML performance were 0.81 (95 % CI 0.76–0.85, I2 78.36 %) and 0.82 (95 % CI 0.77–0.86, I2 72.33 %), respectively. FL perform almost similar with CML and substantially outperform LML across the studies at all prediction time and clinical settings.

Figure 3. Pooled area under the curve (AUC) of federated learning vs centralized machine learning.



Discussion

Principal Results

This study examined the comparison performance between FL and CML or LML approach in developing mortality prediction model. The quality of prediction models developed in each study were assessed in the terms of risk of bias. All included studies had either low risk unclear or high risk of bias. There are two domains were mostly rated as high risk of bias. First is analysis domain, the high risk of bias obtains due to the lack information of missing value. Four included studies had either complete case analysis or not explicitly informed the missing values. Simply excluding enrolled participants with any missing data from the analysis leads to biased predictor–outcome associations and biased model performance. It is suggested to provide either the distributions (percentage, mean, or medians) of the predictors and outcomes between both groups (excluded vs. analyzed participants) or a comparison of the predictor–outcome associations and model predictive performance with and without inclusion of the participants with missing values, otherwise potential risk of bias cannot be easy to judge. Second is outcome domain, the time interval between predictor assessment and outcome occurrence cannot be found in three studies. For both diagnostic and prognostic models, the presence of bias may arise if the outcomes are assessed prematurely, at a time when pertinent outcomes remain undetectable, or when the quantity of outcomes is not representative [25].

In line with previous literature [38, 39], most of the implementations of prediction model development were performed using retrospective cohort, extracting from publicly dataset, rather than using directly follow-up data, only one research employed a prospective cohort dataset to track COVID-19 patient mortality directly. [34]. Current data sources, frequently gathered for objectives distinct from model development, validation, or refinement, exhibit a significant risk of bias attributable to the absence of a standardized protocol. To mitigate this concern, it is recommended to implement a comprehensive quality assessment of the data during the preprocessing phases. [25]. High heterogeneity also identified from the included studies. It can be occurred due to the variability of study populations, study designs or as a result of difference in predictor effects across studies (i.e. due to different measurement methods of predictor).

Various predictors were assessed in accordance to the clinical setting (i.e. demographic, vital signs, lab values, medication). Classifying the patient according to the disease and drug administered for each of them has been proven to provide promising result in improving accuracy, sensitivity, and specificity of FL-based prediction model [28]. However, none of these implementations were taken to production maturity; all were conducted as research studies only, with the exception of Vaid's model, which used their own data [34]. These findings support the fact that FL is still in its infancy and further efforts are needed to move into production phases with FL-based model [15, 19].

Consistent with the prior review [40], prediction result presented in Multimedia Appendix 3, using FL demonstrated the high feasibility and accuracy. The developed FL models achieved AUC value higher than 0.7 and achieve better performance than models only training on one of the private datasets available at each silo. Additionally, the pooled AUC estimate showed slightly differences between FL-based and CML model. The results obtained are yet sufficient to prove the hypothesis of FL feasibility in the field of mortality prediction. This demonstrates that FL is capable of handling large variety of different data types and tasks, namely low-dimensional tabular EHR datasets and high dimensional imaging datasets. In addition, the utilization of real-world cross-silo datasets illustrates that FL possesses the capability to address the complexity and heterogeneity in real-world datasets, thereby underscoring its applicability in practical scenarios and subsequently supporting

human experts. Moreover, the research indicated that models developed through FL exhibit enhanced resilience against privacy threats, such as membership inference attacks, thereby empirically validating the significance of integrating privacy-preserving methodologies for safeguarding patient information [41, 33]. In summary, the FL framework facilitates researchers in executing extensive machine learning investigations and in training more precise models by capitalizing on varied data sources while ensuring the protection of patient privacy. Ultimately, the FL framework therefore represents a promising avenue for fostering secure and private collaboration in machine learning research pertaining to healthcare-related issues.

Strengths and Limitations

Prior to this statement, systematic reviews had already been conducted to assess the quality of machine learning prediction models for different diseases [38]. However, it was not until the end of 2014 that guidelines for good transparency and clarity in this type of study were established through the CHARMS checklist [36]. The CHARMS indicates how to conduct search strategies according to the predictive models to be evaluated (e.g, prognostic models for breast cancer) and, importantly, explains all the information that needs to be extracted from each publication. The information to be extracted is classified into 11 domains: source of data, participants, outcome to be predicted, candidate predictors, sample size, missing data, model development, model performance, model evaluation, results, and interpretation and discussion. Each domain addresses a series of important items for assessing the risk of bias and the applicability of the predictive model to conditions of daily clinical practice. Finally, the recommended procedure when developing or validating a predictive model according to the scientific literature is explained in each domain. Although the CHARMS checklist was introduced in 2014 and numerous systematic reviews of predictive models have been performed thus far [42, 43, 44, 45], to the best of our knowledge, the research typically only engages in a broad descriptive evaluation of the findings from each study included in the review [42, 43, 44, 45] rather than conducting a thorough examination of each predictive model, as our research team does. Another key strength of this systematic review is the quantitative meta-analytical methods used that allow robust conclusions based on cumulative evidence about the performance of FL-based prediction model compare to the classical centralized or local machine learning for the dissemination of research evidence to practitioners. However, we expect future work will further strengthen FL framework review by conducting subgroup analysis in different clinical setting to achieve better enlightenment of the comparison in the terms of discriminant ability as well as privacy-utility-communication trade-offs.

Conclusions

This paper demonstrated that the performance of ML models trained in a federated environment is comparable to those trained on centralized data storage and outperform local machine learning approach. Federated models are not affected by unbalanced data distributions across network nodes. However, training ML models in a federated environment has its cost. The efficiency of model training and inference suffers due to the added complexity of node orchestration, privacy preservation, and extra steps that are not existent in centralized approaches. Experiments showed that model training may take up to 9 times longer, and that inference time may increase by a factor of 40 in comparison to the model trained on centralized data, while the computation round surge as high as 50 to 100 times.

Answering the research question, federated learning approach showed it feasibility to be deployed in developing mortality prediction among hospitalized patient. Federated machine learning has achieved some successes so far, but still faces challenges such as the diversity of data and devices in the FL network and the high cost of computation and communication. This article outlines a number

of insightful pathways that warrant exploration to enhance the efficacy of this technology and to assist prospective researchers in comprehending our current standing with this technology and the essential steps for future advancement.

Acknowledgements

Not applicable.

Data availability

The data sets generated and analyzed during this study will be available from the corresponding author on reasonable request.

Conflicts of Interest

None declared.

Multimedia Appendix 1

PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) checklist.

Multimedia Appendix 2

Search strategy.

Multimedia Appendix 3

Description of the included articles and full details of meta-analyses.

Multimedia Appendix 4

Model summary and approximation formula.

Multimedia Appendix 5

Risk of bias assessment by CHARMS and PROBAST guideline.

References

1. Sarker, I. H. (2021). "Machine Learning: Algorithms, Real-World Applications and Research Directions." SN Computer Science 2(3): 160
2. Sharma, N., et al. (2021). "Machine Learning and Deep Learning Applications-A Vision." Global Transitions Proceedings 2(1): 24-28
3. Zantalis, F., et al. (2019) A Review of Machine Learning and IoT in Smart Transportation. Future Internet 11, DOI: 10.3390/fi11040094
4. Pallathadka, H., et al. (2023). "IMPACT OF MACHINE learning ON Management, healthcare AND AGRICULTURE." Materials Today: Proceedings 80: 2803-2806
5. Erickson, B.J.; Korfiatis, P.; Akkus, Z.; Kline, T.L. Machine learning for medical imaging. Radiographics 2017, 37, 505.
6. Xin, Y., et al. (2018). "Machine Learning and Deep Learning Methods for Cybersecurity." IEEE ACCESS 6: 35365-35381
7. Liakos, K. G., et al. (2018) Machine Learning in Agriculture: A Review. SENSORS 18, DOI: 10.3390/s18082674
8. Larrañaga, P.; Atienza, D.; Diaz-Rozo, J.; Ogbechie, A.; Puerto-Santana, C.; Bielza, C. Industrial Applications of Machine Learning; CRCPress: Boca Raton, FL, USA, 2018.

9. Nagarhalli, T.P.; Vaze, V.; Rana, N.K. Impact of machine learning in natural language processing: A review. In Proceedings of the Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV), IEEE, Tirunelveli, India, 4–6 February 2021; pp. 1529–1534.
10. L'Heureux, A., et al. (2017). "Machine Learning With Big Data: Challenges and Approaches." IEEE ACCESS 5: 7776-7797
11. Zhou, L., et al. (2017). "Machine learning on big data: Opportunities and challenges." Neurocomputing 237: 350-361
12. Parasol, M. (2018). "The impact of China's 2016 Cyber Security Law on foreign technology firms, and on China's big data and Smart City dreams." Computer Law & Security Review 34(1): 67-98
13. Chik, W. B. (2013). "The Singapore Personal Data Protection Act and an assessment of future trends in data privacy reform." Computer Law & Security Review 29(5): 554-575
14. Johnson, A.E., Pollard, T.J., Naumann, T.: Generalizability of predictive models for intensive care unit patients. arXiv preprint arXiv:1812.02275 (2018)
15. Moshawrab, M., et al. (2023). "Reviewing Federated Machine Learning and Its Use in Diseases Prediction." SENSORS 23(4): 2112
16. Maddox, T. M., et al. (2019). "Questions for Artificial Intelligence in Health Care." JAMA 321(1): 31-32
17. Gupta, R., et al. (2021). "Artificial intelligence to deep learning: machine intelligence approach for drug discovery." Molecular Diversity 25(3): 1315-1360
18. Jochems, A., et al. (2017). "Developing and Validating a Survival Prediction Model for NSCLC Patients Through Distributed Learning Across 3 Countries." International Journal of Radiation Oncology*Biology*Physics 99(2): 344-352
19. Benedetto, U., et al. (2022). "Machine learning improves mortality risk prediction after cardiac surgery: Systematic review and meta-analysis." The Journal of Thoracic and Cardiovascular Surgery 163(6): 2075-2087.e2079
20. Diniz, J. M., et al. (2023). "Comparing Decentralized Learning Methods for Health Data Models to Nondecentralized Alternatives: Protocol for a Systematic Review." JMIR Res Protoc 12: e45823
21. Watson, O. J., et al. (2022). "Global impact of the first year of COVID-19 vaccination: a mathematical modelling study." The Lancet Infectious Diseases 22(9): 1293-1302
22. Ku, E., et al. (2022). "Comparison of 2021 CKD-EPI Equations for Estimating Racial Differences in Preemptive Waitlisting for Kidney Transplantation." CLINICAL JOURNAL OF THE AMERICAN SOCIETY OF NEPHROLOGY 17(10)
23. Page, M. J., et al. (2021). "The PRISMA 2020 statement: an updated guideline for reporting systematic reviews." Bmj 372: n71
24. Debray, T. P., et al. (2017). "A guide to systematic review and meta-analysis of prediction model performance." Bmj 356: i6460
25. Moons, K. G. M., et al. (2019). "PROBAST: A Tool to Assess Risk of Bias and Applicability of Prediction Model Studies: Explanation and Elaboration." Ann Intern Med 170(1): W1-w33
26. Fernandez-Felix, B. M., et al. (2023). "CHARMS and PROBAST at your fingertips: a template for data extraction and risk of bias assessment in systematic reviews of predictive models." BMC Medical Research Methodology 23(1): 44
27. Budrionis, A., et al. (2021). "Benchmarking PySyft Federated Learning Framework on MIMIC-III Dataset." IEEE ACCESS 9: 116869-116878
28. Huang, L., et al. (2019). "Patient clustering improves efficiency of federated machine learning to predict mortality and hospital stay time using distributed electronic medical records." JOURNAL OF BIOMEDICAL INFORMATICS 99: 103291

29. Kerkouche, R., et al. (2021). Privacy-preserving and bandwidth-efficient federated learning: an application to in-hospital mortality prediction. Proceedings of the Conference on Health, Inference, and Learning. Virtual Event, USA, Association for Computing Machinery: 25–35
30. Li, S., et al. (2023). "FedScore: A privacy-preserving framework for federated scoring system development." JOURNAL OF BIOMEDICAL INFORMATICS 146: 104485
31. Pfitzner, B., et al. (2024). Differentially-Private Federated Learning with Non-IID Data for Surgical Risk Prediction. 2024 IEEE First International Conference on Artificial Intelligence for Medicine, Health and Care (AIMHC)
32. Randl, K., et al. (2023). Early prediction of the risk of ICU mortality with Deep Federated Learning. 2023 IEEE 36th International Symposium on Computer-Based Medical Systems (CBMS)
33. Shiri, I., et al. (2024). "Differential privacy preserved federated learning for prognostic modeling in COVID-19 patients using large multi-institutional chest CT dataset." Med Phys
34. Vaid, A., et al. (2021). "Federated Learning of Electronic Health Records to Improve Mortality Prediction in Hospitalized Patients With COVID-19: Machine Learning Approach." JMIR Med Inform 9(1): e24207
35. Zhou, J., et al. (2024). "Federated-learning-based prognosis assessment model for acute pulmonary thromboembolism." BMC Med Inform Decis Mak 24(1): 141
36. Palazón-Bru, A., et al. (2020). "A general presentation on how to carry out a CHARMS analysis for prognostic multivariate models." Stat Med 39(23): 3207-3225
37. Awad, A., et al. (2019). "Predicting hospital mortality for intensive care unit patients: Time-series analysis." HEALTH INFORMATICS JOURNAL 26(2): 1043-1059
38. Andaur Navarro, C., et al. (2021). "Risk of bias in studies on prediction models developed using supervised machine learning techniques: systematic review." BMJ 375(8311)
39. Frondelius, T., et al. (2024). "Early prediction of ventilator-associated pneumonia with machine learning models: A systematic review and meta-analysis of prediction model performance." Eur J Intern Med 121: 76-87
40. Dang, T. K., et al. (2020). Building ICU In-hospital Mortality Prediction Model with Federated Learning. Federated Learning: Privacy and Incentive. Q. Yang, L. Fan and H. Yu. Cham, Springer International Publishing: 255-268
41. Fang, C., et al. (2024). "Decentralised, collaborative, and privacy-preserving machine learning for multi-hospital data." eBioMedicine 101: 105006
42. Collins, G. S., et al. (2013). "A systematic review finds prediction models for chronic kidney disease were poorly reported and often developed using inappropriate methods." J Clin Epidemiol 66(3): 268-277
43. Collins, G. S., et al. (2011). "Developing risk prediction models for type 2 diabetes: a systematic review of methodology and reporting." BMC Medicine 9(1): 103
44. Feng, Q., et al. (2019). "Prognostic Models for Predicting Overall Survival in Patients with Primary Gastric Cancer: A Systematic Review." Biomed Res Int 2019: 5634598
45. Lindroth, H., et al. (2018). "Systematic review of prediction models for delirium in the older adult inpatient." BMJ Open 8(4): e019223

Supplementary Files

Multimedia Appendixes

Untitled.
URL: http://asset.jmir.pub/assets/db7b85dffe09633f1ac89eb67285726b.docx

Untitled.
URL: http://asset.jmir.pub/assets/303d4398b0c27e77e9168d5ef02be464.docx

Untitled.
URL: http://asset.jmir.pub/assets/3c906fe379e7525f8df461a1c071e919.docx

Untitled.
URL: http://asset.jmir.pub/assets/74690b39d2e9e682c3b7ea23d31b767f.docx

Untitled.
URL: http://asset.jmir.pub/assets/049f445503bf7725b185896f46678b77.docx