Federated Learning for Electronic Health Records

TRUNG KIEN DANG and XIANG LAN, Saw Swee Hock School of Public Health, National University of Singapore, Singapore
JIANSHU WENG, AI Singapore, Singapore
MENGLING FENG, Institute of Data Science & Saw Swee Hock School of Public Health, National University of Singapore, Singapore

In data-driven medical research, multi-center studies have long been preferred over single-center ones due to a single institute sometimes not having enough data to obtain sufficient statistical power for certain hypothesis testings as well as predictive and subgroup studies. The wide adoption of electronic health records (EHRs) has made multi-institutional collaboration much more feasible. However, concerns over infrastructures, regulations, privacy, and data standardization present a challenge to data sharing across healthcare institutions. Federated Learning (FL), which allows multiple sites to collaboratively train a global model without directly sharing data, has become a promising paradigm to break the data isolation. In this study, we surveyed existing works on FL applications in EHRs and evaluated the performance of current state-of-the-art FL algorithms on two EHR machine learning tasks of significant clinical importance on a real world multi-center EHR dataset.

CCS Concepts: • **Applied computing** → *Health informatics;*

Additional Key Words and Phrases: Federated learning, electronic health records, healthcare, neural networks

ACM Reference format:
Trung Kien Dang, Xiang Lan, Jianshu Weng, and Mengling Feng. 2022. Federated Learning for Electronic Health Records. *ACM Trans. Intell. Syst. Technol.* 13, 5, Article 72 (June 2022), 17 pages.
https://doi.org/10.1145/3514500

1 INTRODUCTION

The broad adoption of **electronic health records (EHRs)** presents opportunities for collaboration among hospitals. For medical research, multi-center studies have long been considered superior to single-center ones. The larger combined cohort allows for certain hypothesis testings and subgroup analyses that are often not possible in a single-center setting due to inadequate statistical power [46, 70]. In machine learning, multi-center datasets could lead to more robust models.

Trung Kien Dang and Xiang Lan contributed equally to this research.
This project is supported by the National Research Foundation, Singapore under its AI Singapore Programme (AISG Awards No. AISG-100E-2020-055 and No. AISG-GC-2019-002A) and also the NMRC HSRG MOH-000030-00. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of National Research Foundation, Singapore.
Authors' addresses: T. K. Dang and X. Lan, Saw Swee Hock School of Public Health, National University of Singapore, Tahir Foundation Building, 12 Science Drive 2, #10-01, 117549, Singapore; emails: kiendang@u.nus.edu, ephlanx@nus.edu.sg; J. Weng, AI Singapore, innovation 4.0, 3 Research Link, #02-05, 117602, Singapore; email: js@aisingapore.org; M. Feng (corresponding author), Institute of Data Science, innovation 4.0, 3 Research Link, #04-06, 117602, Singapore & Saw Swee Hock School of Public Health, National University of Singapore, Tahir Foundation Building, 12 Science Drive 2, #10-01, 117549, Singapore; email: ephfm@nus.edu.sg.

2157-6904/2022/06-ART72
https://doi.org/10.1145/3514500

A model trained only on single-center data is prone to poor generalizability, i.e., it may only perform well on data from the same hospital that provided the training data but do poorly when applied to data from others [25]. This is potentially due to differences among hospitals in medical practices, patient demographics, genotypes and phenotypes, as well as variations in software, hardware, and protocols used for data collection. Environmental, social, political, and cultural variations may also play a part. Multi-center datasets may enable models to capture and adapt to the heterogeneity caused by these factors and thus improve their generalizability [25, 63, 81]. In addition, simply by collecting data from several sources, studies end up with a larger dataset for training, which reduces the expected generalization error of the model [24].

Despite the benefits, in reality, conducting machine learning on EHRs across multiple sources faces several tough challenges. The traditional approach of centralized model training is to gather datasets from different silos and store them in a centralized data warehouse so that machine learning models could be trained on the combined dataset. In practice, such collaborative EHR repositories have been established by healthcare organizations who were willing to pool their data together to conduct their research on a larger scale [49, 69]. However, these efforts faced multiple challenges regarding logistics, infrastructures, regulations, privacy and data standardization [12, 73, 75]. Central data repositories increase the risk of data security and privacy compromises. Examples include data leakage due to an increase in the number of parties with access to the data as well as subject re-identification due to linkage across multiple data sources [21]. Moreover, EHRs are subject to a set of rigorous regulations regarding accessing, analyzing and sharing personal health information [3, 27, 79]. Additionally, each hospital also imposes its internal policies on the matter. Significant efforts are required to ensure compliance with these regulations and policies. Last but not least, the high cost to set up and maintain the infrastructure for centralized data storage presents yet another roadblock against collaborative machine learning based on data centralization.

Given these challenges associated with centralized learning, a method that enables collaborative model training to occur in a decentralized manner, without the need for aggregating all data in one place, would make machine learning on EHRs across multiple centers much more feasible. **Federated Learning (FL)** is an emerging paradigm that enables building machine learning models collaboratively using decentralized data. It was originally proposed by Google for the use case of Gboard query suggestion [43]. The project involved developing a language model over hundreds of thousands of mobile devices for keyboard autosuggestion, predicting the most likely words that a user would type next. Each participating device trained a separate local model using only their own local dataset. Local models were then sent to a central coordinator where they were aggregated into a global model to be sent back to each participant, either for inference or further training. The main purpose of FL is to enable participants to collaborate and produce a better model than they could on their own without compromising data privacy. It achieves this by requiring participants to share only model parameters, not data.

Given these characteristics, FL has the potential to help facilitate collaboration on data-driven research on EHRs across multiple institutions while preserving data privacy. However, that FL, or horizontal FL to be specific, requires data from participating parties to be of the same format might present a challenge. Recently there have been increasing efforts by healthcare institutions towards data harmonization. More and more organizations are adopting a common data model, such as i2b2 [55], PCORnet [22], and OMOP CDM [80], for their EHRs. These organizations are thus well-positioned to employ FL to facilitate collaborations with others who utilize the same data model [17].

In this study, we give a brief survey of existing applications of FL on EHR data. We then provide an overview of common FL algorithms and evaluate their performance on two EHR machine learning tasks, in-hospital mortality prediction and **acute kidney injury (AKI)** prediction in the

intensive care unit (ICU). These two tasks have significant clinical importance and have been shown to greatly benefit from machine learning. In literature there exist various state-of-the-art FL algorithms that achieve good results on general domain or benchmark datasets. However, there has not been a study that evaluates how well they perform in the context of EHRs. Related to this work, Xu et al. [83] published a survey that summarized the progress and challenges of FL as well as gave an overview of current applications and potential opportunities of FL in healthcare. In addition, there exist many studies that successfully applied FL to outcome prediction using EHRs [32, 59, 66, 77]. To the best of our knowledge, ours is the first one that aims to systematically compare the performance of several state-of-the-art FL algorithms on an EHR machine learning task.

Among state-of-the-art FL methods, the most well-known algorithm is Federated Averaging, also known as FedAvg [50]. FedAvg builds a shared global model by periodically averaging the weights of the locally evolving models. Despite being the standard federated optimization method, it suffers from slow convergence or even divergence in some situations when data distribution differs among clients. Multiple variants of FedAvg have been shown to improve the algorithm convergence behavior [30, 31, 38, 45, 62]. We conducted a previous pilot study investigating the performance of FedAvg and FedProx [45], another popular FL method, on predicting in-hospital mortality [14]. In this study, we extended our earlier work by inspecting a more comprehensive set of FL algorithms and evaluating them on an additional task of AKI prediction.

2 EXISTING APPLICATIONS OF FEDERATED LEARNING ON ELECTRONIC HEALTH RECORDS

Wide adoptions of EHRs among healthcare institutions have given rise to various studies on applying machine learning to biomedical research [52]. However, conducting machine learning on EHRs is not without challenges. Among those, lack of data and poor model generalizability are two major problems. For research projects involving rare diseases and conditions, small hospitals may not have enough data for machine learning models to learn meaningful patterns. In addition, as mentioned above, machine learning models trained on data obtained from a single source may not generalize well and thus perform poorly when applied to a different context. FL is becoming a promising approach to mitigate these shortcomings. It allows training a global model on a larger and more diverse set of EHRs from multiple institutions while keeping the data locally, thereby preserving privacy and enhancing the model's external validity. Several studies have looked into the effectiveness of solving healthcare problems in an FL setting using EHRs. They can be summarized into two categories, predictive modeling and representation learning.

Many studies have achieved success in applying FL to predictive modeling on EHRs. Sharma et al. [66] proposed an FL framework to predict in-hospital mortality for patients in the ICU. Results showed performance obtained by models trained in an FL setting to be on par with those trained in a centralized manner. Vaid et al. [77] achieved an improvement in predicting 7-day mortality for hospitalized COVID-19 patients by employing FL to utilize data from 5 different hospitals. To predict preterm birth in the context of distributed EHRs, Boughorbel et al. [8] presented a **federated uncertainty-aware learning algorithm (FUALA)** based on FedAvg. FUALA is capable of dynamically adjusting the aggregation model weights by taking into account each model's uncertainty, thus reducing the adverse effects of models with high uncertainties. Brisimi et al. [10] proposed an iterative **cluster Primal-Dual Splitting (cPDS)** algorithm for solving the large-scale soft-margin l1-regularized sparse Support Vector Machine to predict hospitalizations due to cardiac events in FL settings. Huang el al. [33] proposed an FL algorithm called LoAdaBoost to predict the mortality of patients admitted to the ICU based on drugs prescribed during the first 48 h of their ICU stay. Pfohl et al. [59] comprehensively studied the efficacy of FL and differential privacy

versus centralized training in predicting prolonged length of stay and in-hospital mortality across thirty-one hospitals. Grama et al. [28] evaluated the performance of different robust FL aggregation methods on two disease prediction tasks, diabetes mellitus onset prediction and heart failure prediction. Results showed that **adaptive federated averaging (AFA)** [54] not only performs well but also is robust against malicious or faulty clients. Tan et al. [72] proposed a tree-based FL method for treatment effect estimation. The method was used to study the effect of oxygen saturation on hospital mortality among ICU patients with respiratory diseases. Tuladhar et al. [76] presented an ensemble approach to distributed learning of machine learning models for rare disease detection. In this approach, inference is done by ensembling predictions provided by local models instead of aggregating their weights to produce a single global model. Xue et al. [85] introduced a federated reinforcement learning system that employs **Double Deep Q-Network (DDQN)** to provide supports for personalized clinical decisions. The system utilizes data from smart devices at the edge as well as **electronic medical records (EMRs)**.

FL has also been applied to representation learning in the context of EHRs. Liu et al. [47] proposed a two-stage federated **natural language processing (NLP)** method for phenotyping and patient representation learning. The first stage constructs a representation of patient data using medical notes from multiple hospitals without sharing the notes. The learned presentation is not constrained to any specific medical task. The second stage builds a machine learning model for a specific phenotyping task based on relevant features extracted from representations learned in the first stage. Lee et al. [44] and Xu et al. [84] presented two federated patient hashing frameworks for patient similarity learning. The model learns context-specific hash codes to represent patients across multiple hospitals. The learned hash codes are then used to calculate similarities among patients. Ultimately, the model can match patients with high similarity among multiple hospitals. Lu et al. [48] proposed an efficient decentralized FL approach to extract latent features from patient data. Kim et al. [41] proposed a tensor factorization method that generates meaningful clinical concepts (phenotypes) from a large volume of EHRs. Vepakomma et al. [78] introduced three configurations of a distributed deep learning method called Split Learning [29], which differs from conventional FL in that it does not require participants to share the weights of the entire locally trained model. This leads to improved data privacy and security. Huang et al. [32] proposed a method called **community-based federated learning (CBFL)**, which clusters the distributed patient data into clinically meaningful groups that share similar characteristics, such as drug features and diagnoses, while simultaneously training one model for each group. The method achieved good results on predicting mortality and length of stay.

3 EVALUATION OF CURRENT WELL-KNOWN FL ALGORITHMS ON EHRS

This section provides an overview of common FL algorithms that have been shown to work well outside of healthcare domain. Their performance was then evaluated on two machine learning tasks in the ICU, in-hospital mortality prediction and AKI prediction, using a dataset containing EHRs from multiple ICUs.

3.1 Overview of Common FL Algorithms

In general, FL involves each individual participants training local models on their local dataset alone and then exchanging model parameters, e.g., the weights and or gradients, at some frequency. There is no exchange of data among different participants. The local model parameters are then aggregated to generate a global model. Aggregation can be conducted with or without the coordination of a central party. Different FL algorithms vary in how the aggregation steps or the local update steps are performed. Among those, FedAvg [50], is the most well-known. FedAvg aims to

optimize the following objective:

$$\min_{\mathrm{w}} \left(F(\mathrm{w}) = \sum_{k=1}^{K} p_k F_k(\mathrm{w}) \right), \tag{1}$$

where N is the number of participants and p_k is the weight of participant k and $\sum_{i=k}^{N} p_k = 1$. p_k is usually proportional to the size of each participant dataset. $F_k(\cdot)$ is the local objective function.

At each communication round t, a global model with weights w_t is sent to all K participants. Each participant k performs local training for E epochs, producing a new local model with weights w_{t+1}^k. Each participant then sends their newly learned local model weights to a central server where they are aggregated to obtain a new global model with updated weights w_{t+1} equal to the weighted average of all local models:

$$w_{t+1} = \sum_{k=1}^{K} p_k w_{t+1}^k. \tag{2}$$

FedAvg performs well in the case of homogeneity, where all local datasets are **identically and independently distributed (IID)**. In the presence of statistical heterogeneity where data are not identically and independently distributed (non-IID) across participants, the global model might perform poorly or not even converge. A number of different approaches have been proposed to counter this problem and improve the convergence rate and performance of FL for non-IID datasets.

FedProx [45] and SCAFFOLD [38] aim to improve the convergence rate in FedAvg by correcting *client drift*, a phenomenon where client heterogeneity causes a drift in the local updates in each round of local training, resulting in slow convergence. FedProx introduces a *proximal term* that restricts the local updates to be closer to the latest global update. Instead of optimizing $F_k(\cdot)$, each participant now optimizes the local objective:

$$h_k(\mathrm{w}, \mathrm{w}^t) = F_k(\mathrm{w}) + \frac{\mu}{2} \left\| \mathrm{w} - \mathrm{w}^{\mathrm{t}} \right\|^2. \tag{3}$$

SCAFFOLD works by measuring the amount of drift caused by each client in each round and then adjusts their local update accordingly. How much a client drifts is measured by the difference in the direction of the global update versus the direction of the client local update.

Instead of controlling local training, several FL algorithms tackle the slow convergence problem by experimenting with server optimization. The global model update step specified in Equation (2) can be rewritten as

$$w = w - \Delta w, \tag{4}$$

where $\Delta w = \sum_{k=1}^{K} p_k \Delta w_k$ and Δw_k is the weight updates from client k,

$$\Delta w_k = w_k^{t+1} - w^t. \tag{5}$$

$w = w - \Delta w$ has the same form as a gradient-based optimization step where Δw acts as a *pseudo-gradient*. Reddi et al. [62] formalized this as a server optimization step that optimizes the model from a global perspective, in addition to the client optimization step 5 that aims to optimize the model from a local perspective. Their proposed FL algorithms FedAdagrad, FedAdam, and FedYogi employ adaptive server optimization by applying adaptive optimization methods Adagrad, Adam, and Yogi in the server optimization step. FedAvgM [30, 31] is another algorithm that uses adaptive server optimization, by adding momentum to the server optimization step , computing $w = w - v$ where $v = \beta v + \Delta w$.

3.2 Experiments

We evaluated the performance of well-known FL algorithms, FedAvg, FedProx, FedAvgM, FedAdagrad, FedAdam, and FedYogi, on two common and clinically crucial machine learning tasks in the ICU, in-hospital mortality prediction and AKI prediction. Their results were compared against those obtained from local learning, centralized learning and two non-FL methods that also enable collaborative model training without data sharing, namely, IIL and **cyclic institutional incremental learning (CIIL)** [11, 67, 68]. In IIL, each party trains the model on their local dataset then passes the model to the next one until all parties have trained the model. CIIL repeats the same process over multiple rounds, but fixes the number of training epochs carried out by each party at each round. The data for both tasks come from the eICU dataset [60], which collected EHRs from more than 200 hospitals and over 139,000 patients across the United States admitted to the ICU in 2014 and 2015. The dataset contains a wide range of data, including demographics, medication, diagnoses, procedures, timestamped vital signs, and lab test results.

For each task, several hospitals in the eICU database were selected as participants. The extracted data were split into a train, validation, and test set for each of the hospitals, each taking up 80%, 10%, and 10% of the whole population, respectively. In the local training setting, a separate model was trained for each hospital using only their own local data. The training was done over a number of epochs, and for each hospital, the model that gave the best performance on the validation set in terms of **Area under the ROC Curve (AUC-ROC)** became the final model for evaluation.

In the centralized setting, the train, validation, and test sets from all participating hospitals were concatenated to produce a single train, validation, and test set. A single model was then trained on the combined training set. Like in the local setting, training was conducted for several epochs and the best model was picked based on the AUC-ROC score on the combined validation set.

In the IIL and CIIL settings, since there was no global aggregated validation set due to no data sharing among participants, the model produced by the last party that conducted the training was selected as the final model.

In the FL setting, training was done over several communication rounds. Similar to the IIL and CIIL settings, since the central server that coordinated the training and carried out the global model aggregation process did not have access to a global validation set, the final model was the one obtained after all the communication rounds had finished.

Performance among the methods was compared based on global test scores. The metrics used are AUC-ROC and **Area under the Precision-Recall Curve (AUC-PR)**. Delong's method [15] and logit method [9] were employed to compute 95% confidence intervals for AUC-ROC and AUC-PR, respectively.

3.2.1 In-hospital Mortality Prediction. In this experiment, we investigated the performance of FL algorithms on predicting a patient's in-hospital mortality based on data collected during the first 24 h of their ICU stay. This is a crucial task in clinical setting. When a patient is admitted to the ICU, predicting their mortality, either at the end of the ICU stay, hospital stay, or within a fixed period, e.g., 28 days, one month, or three months, provides a proxy for the severity of their condition and helps healthcare providers plan treatment pathways and allocate resources more effectively. There exist several works on successfully applying machine learning to predict in-hospital mortality [6, 61, 82].

Data. The same data extraction process in References [14, 37] was employed. For each hospital in the entire eICU dataset, we extracted a cohort of patients age 16 and above in their first ICU stay who had their in-hospital mortality status recorded. Patients without an APACHE IVa score were excluded. This criterion serves as a proxy for identifying patients with insufficient data or those who were only in the database for administration purpose. Twenty hospitals with the largest

cohorts were then selected as participants in the study. The combined cohort contains 87,003 ICU stays.

For each patient, data within 24 h from ICU admission were extracted. The set of features includes

- demographic information: gender, age, and ethnicity,
- the first and last results of the following laboratory tests: PaO2, PaCO2, PaO2/FiO2 ratio, pH, base excess, Albumin, the significant band of arterial blood gas, HCO3, Bilirubin, **Blood Urea Nitrogen (BUN)**, Calcium, Creatinine, Glucose, Hematocrit, Hemoglobin, **international normalized ratio (INR)**, Lactate, Platelet, Potassium, Sodium, white blood cell count,
- the first and last as well as the minimum and maximum measurements of the following vital signs: heart rate, systolic blood pressure, mean blood pressure, respiratory rate, temperature (Celcius), SpO2, **Glasgow Coma Scale (GCS)**,
- total urine output,
- whether the hospital admission was for an elective surgery.

A total of 82 covariates were obtained.

Methods. A neural network consisting of two fully connected hidden layers with ReLU activation function and L2 normalization was used. The first hidden layer contains 100 nodes and the second 50. In the local and centralized settings, the model was trained for 90 epochs. In FL settings, the training took place over 30 communication rounds, with each hospital training the model locally for ten epochs each round.

3.2.2 AKI Prediction. The purpose of this experiment was to evaluate the performance of FL algorithms on predicting the risk of a patient developing AKI within the next hour based on data collected during the previous 7 h. AKI is a sudden onset of renal damage or kidney failure that happens within a few hours or a few days and occurs in at least 5% of hospitalized patients [16]. AKI can affect other organs such as lungs, heart, and brain. It significantly increases hospitalization cost as well as mortality risk [13]. A timely detection of AKI could prevent patients from developing chronic kidney disease [39, 71]. There have been several studies that show strong performance of machine learning models in predicting AKI [26, 53, 58].

Data. We followed the same data extraction process in Reference [16]. The RIFLE criteria [7] were used to define AKI. Specifically, a patient at time t will be labeled as suffering from AKI if their urine output is less than 0.5 ml/kg/h for $t >= 6$. The cohort exclusion criteria include (1) patients who were under 16 years old or stayed in the ICU for less than 12 h and (2) patients whose data for the selected variables were not recorded at least once during their ICU stay. A total of 10,967 patients in 168 hospitals remained after the filtering. The top 75% hospitals with the most number of patients were selected to participate in the study. The final cohort contains 28 hospitals with a total of 6,641 patients.

For each patient, we extracted data in 7-h sliding windows. The full set of covariates includes

- demographic information: age and gender,
- the minimum and maximum values as well as the range (the difference between the maximum and minimum values) of the following vital signs: heart rate, respiratory rate, mean blood pressure,
- the minimum and maximum values as well as the range of the following lab measurements: SpO2/SaO2, pH, Potassium, Calcium, Glucose, Sodium, HCO3, Hemoglobin, white blood cell count, Platelet count, Urea Nitrogen, Creatinine, GCS,

Table 1. Global Test Performance on the In-hospital Mortality Prediction Task

Method	AUC-ROC (95% CI)	AUC-PR (95% CI)
Local	0.833 (0.808–0.859)	0.472 (0.373–0.480)
Centralized	0.918 (0.907–0.929)	0.668 (0.633–0.701)
IIL	0.877 (0.857–0.897)	0.505 (0.451–0.559)
CIIL	0.828 (0.803–0.852)	0.426 (0.373–0.480)
FedAvg	0.901 (0.882–0.921)	0.638 (0.584–0.688)
FedProx	0.895 (0.877–0.914)	0.577 (0.523–0.630)
FedAvgM	0.906 (0.888–0.925)	0.645 (0.591–0.695)
FedAdam	0.890 (0.870–0.911)	0.578 (0.524–0.631)
FedAdagrad	0.893 (0.873–0.913)	0.596 (0.543–0.649)
FedYogi	0.895 (0.875–0.915)	0.594 (0.539–0.646)

Table 2. Global Test Performance on the AKI Prediction Task

Method	AUC-ROC (95% CI)	AUC-PR (95% CI)
Local	0.709 (0.697–0.722)	0.748 (0.734–0.761)
Centralized	0.735 (0.724–0.747)	0.783 (0.770–0.796)
IIL	0.664 (0.652–0.677)	0.723 (0.709–0.737)
CIIL	0.712 (0.70–0.724)	0.764 (0.750–0.777)
FedAvg	0.724 (0.712–0.736)	0.770 (0.757–0.783)
FedProx	0.691 (0.679–0.703)	0.740 (0.726–0.754)
FedAvgM	0.725 (0.713–0.736)	0.775 (0.762–0.788)
FedAdam	0.716 (0.704–0.728)	0.760 (0.746–0.773)
FedAdagrad	0.720 (0.708–0.732)	0.767 (0.753–0.780)
FedYogi	0.732 (0.720–0.743)	0.773 (0.760–0.786)

- interventions: use of vasoactive medications, use of sedative medications, and use of mechanical ventilation.
- total urine output.

A total of 22 covariates were obtained.

Methods. Similar to the previous task, a fully connected neural network consisting of two hidden layers with ReLU activation function and L2 normalization was used. However, here each of the two hidden layers contains 512 nodes instead of 100 and 50. In the local and centralized settings, the model was trained for 30 epochs. In FL settings, training took place over four communication rounds. Each hospital trained a local model for 10 local epochs during the first round and 5 local epochs during each subsequent round.

3.3 Results and Discussion

Global test performance in terms of AUC-ROC and AUC-PR obtained with each method is shown in Table 1 for in-hospital mortality prediction and Table 2 for AKI prediction. Comparison of ROC curves obtained with FL methods versus centralized and local training is visualized in Figures 1 and 2. Similarly, Figures 3 and 4 in Appendix A show comparison of ROC curves obtained with FL

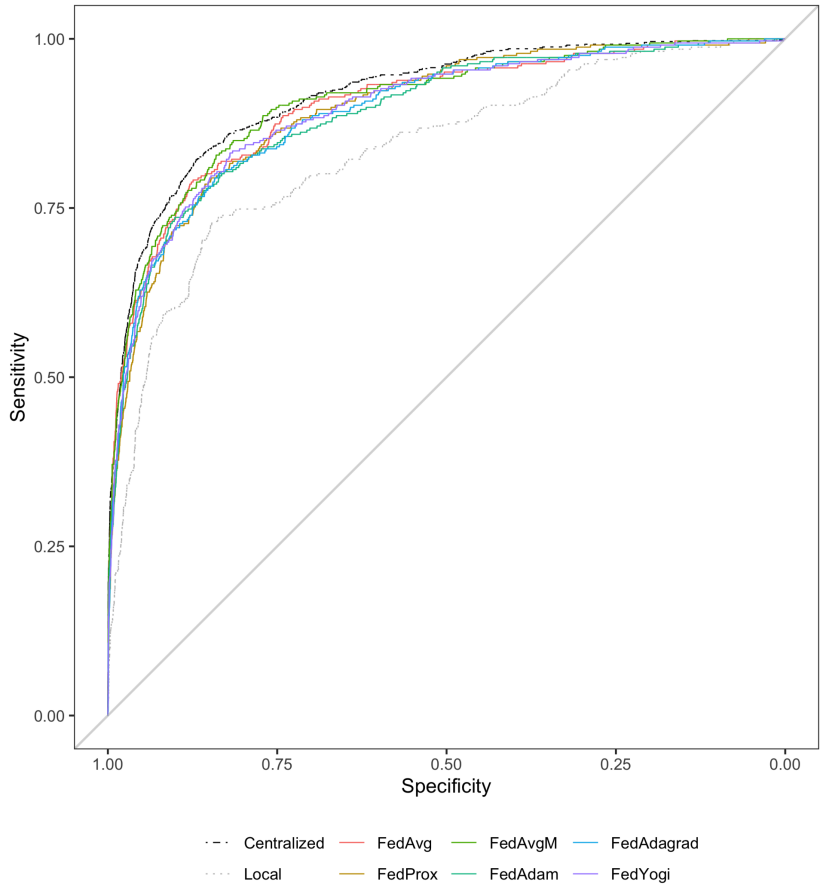

Fig. 1. Comparison of ROC curves obtained with Local, Centralized, and FL methods for in-hospital mortality prediction.

methods compared to those obtained with CIIL and IIL. In both tasks, all FL methods outperform local training in either metric with the exception of FedProx in predicting AKI. In particular, for mortality prediction, all FL methods perform significantly better than local training. In comparison with IIL and CIIL, for mortality prediction, all FL methods achieve better results. For AKI prediction, the same is true for most FL methods. Only exceptions are FedProx, which obtains worse AUC-ROC and AUC-PR than both IIL and CIIL, and FedAdam, whose AUC-PR is slightly lower than that of CIIL. Overall, for both tasks, FL methods enjoy improvement over IIL and CIIL. This is unsurprising given that IIL is known to suffer from catastrophic forgetting [23, 42, 68] while it is non-trivial to obtain optimal results with CIIL due to its instability [68]. Results obtained by FL are also comparable to centralized learning, with the best FL method in each task achieving AUC-ROC within 0.01 of the global AUC-ROC for centralized learning in terms of point estimates. FedAvg and FedAvgM perform consistently well and are among the top three FL methods with the highest global AUC-ROCs and AUC-PRs in either task, only behind FedYogi in AKI prediction. In both cases, FedAvgM obtained slightly better results than FedAvg. However, FedProx achieved the lowest scores in both mortality and AKI prediction.

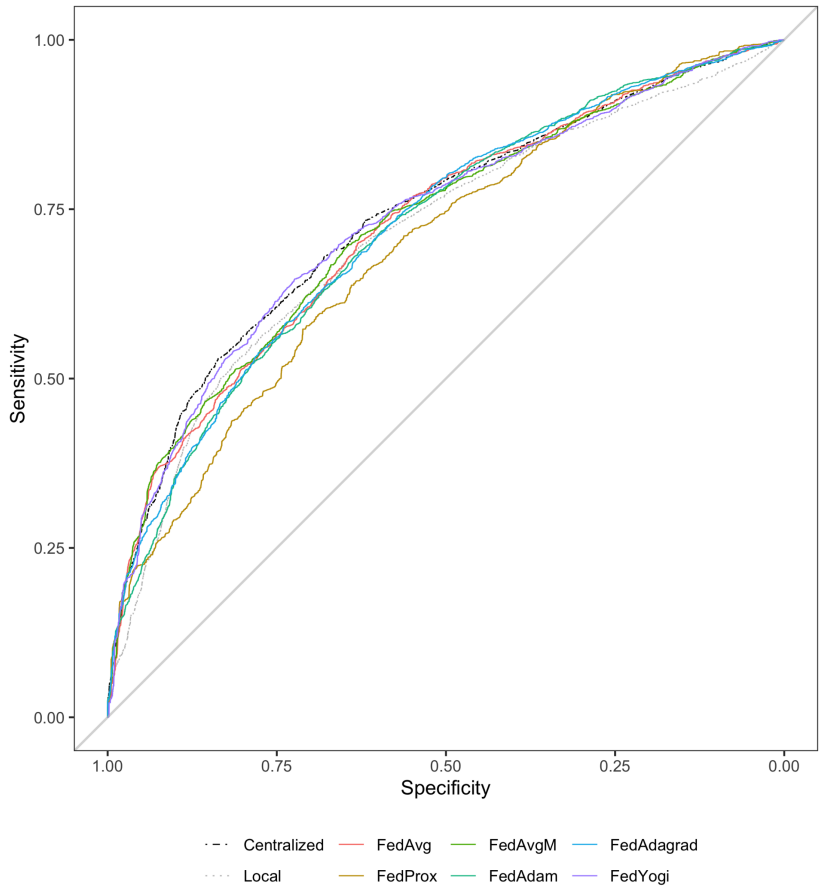

Fig. 2. Comparison of ROC curves obtained with Local, Centralized, and FL methods for AKI prediction.

Results strongly favor FL as a viable strategy for facilitating collaboration among organizations in clinical research. Even though performance does not vary much among the different FL methods in our experiments, it is observed that simple FL algorithms, namely, FedAvg and FedAvgM, perform slightly better than FedProx, FedAdam and FedAdagrad. It has been shown that FedProx works well in the presence of heavy data heterogeneity [45]. In our dataset, all hospitals are located in the United States and therefore expected to experience consistencies in clinical practices and patient demographics. Furthermore, they all participated in the Philips eICU program, which guarantees a certain degree of data standardization. Thus, the differences in data distribution among them are not significant enough to benefit from FedProx. Plus, the total number of participants is relatively small compared to FL in an IoT setting with a large number of participating devices where FedProx usually shines [45]. Data homogeneity might also contribute to the lack of performance gain in FedAdam and FedAdagrad compared to FedAvg. In addition, both the tasks of predicting mortality and predicting AKI, similar to most machine learning tasks on tabular EHR data, only require the use of feed forward fully connected neural networks with a small number of layers, which might not see considerable performance gain through the use of Adam and Adagrad.

4 CONCLUSION

This study gave a brief survey on applications of FL on EHR data and then evaluated the performance of multiple common FL algorithms on two typical EHR machine learning tasks, in-hospital mortality prediction and AKI prediction. FL shows notable improvement compared to local training and performs close to centralized learning. This is promising for organizations that seek to collaborate with others in data-driven clinical research using EHRs. FL could help them build better machine learning models than individually using only their own local data while preserving data privacy without compromising on model performance.

Our results also suggest that simple FL algorithms FedAvg and FedAvgM work particularly well for machine learning tasks on tabular EHR compared to more complex methods such as FedProx, FedAdam and FedAdagrad. Data homogeneity due to the fact that all local datasets in our experiments come from hospitals located in the U.S. and thus share certain characteristics might contribute to this finding. This is one limitation in our study that we aim to overcome in future works. We plan to expand the pool of participants in our experiments to include datasets from ICUs in Europe [34, 74], Australia, and New Zealand [70] and investigate how FL performs on more heterogeneous EHR data as a future research direction.

In addition, we plan to validate our results on more recent data. We are currently working with public hospitals in Singapore to establish a federated data network whose participants adopt the OMOP Common Data Model [65, 80] for their EHRs. Once the network is set up, we would replicate our study on this more up to date dataset and also expand it to cover more tasks other than mortality and AKI prediction to improve the generalizability of our findings. Another direction would be to evaluate the performance of different FL methods on other data modals in healthcare such as time series, digital signals [2] and medical imaging [35, 36, 56], those that require more complex model architectures.

It is important to note that even though FL aims to preserve data privacy by sharing model parameters instead of data during training, it by itself does not mitigate all data privacy and security concerns. Studies showed that it is possible to make inference about the raw data by examining model parameter updates [5]. There exists methods that add extra security measures on top of FL to counter this [1, 4, 5, 40], namely, **Differential Privacy (DF)** [18–20], **Homomorphic Encryption (HE)** [64] and **Secure Multi-Party Computation (SMC)** [51, 86]. They enhance FL data security and privacy at the cost of communication efficiency and model performance. In particular, by adding noise to client training data, DF offers improvement in data privacy but also results in a decrease in model accuracy. HE ensures that only encrypted model parameters are exchanged. This provides data protection but also imposes a penalty on model performance [57]. SMC preserves knowledge of client inputs but is computationally intensive and requires extensive communication among parties. Further research is needed to understand the privacy-accuracy trade-offs of combining these methods with different FL algorithms in the context of EHR.

APPENDIX

A COMPARISON OF ROC CURVES OBTAINED BY FL METHODS VERSUS IIL AND CIIL

Figures 3 and 4 show comparison of ROC curves obtained with IIL, CIIL, and FL methods for in-hospital mortality prediction and AKI prediction, respectively.

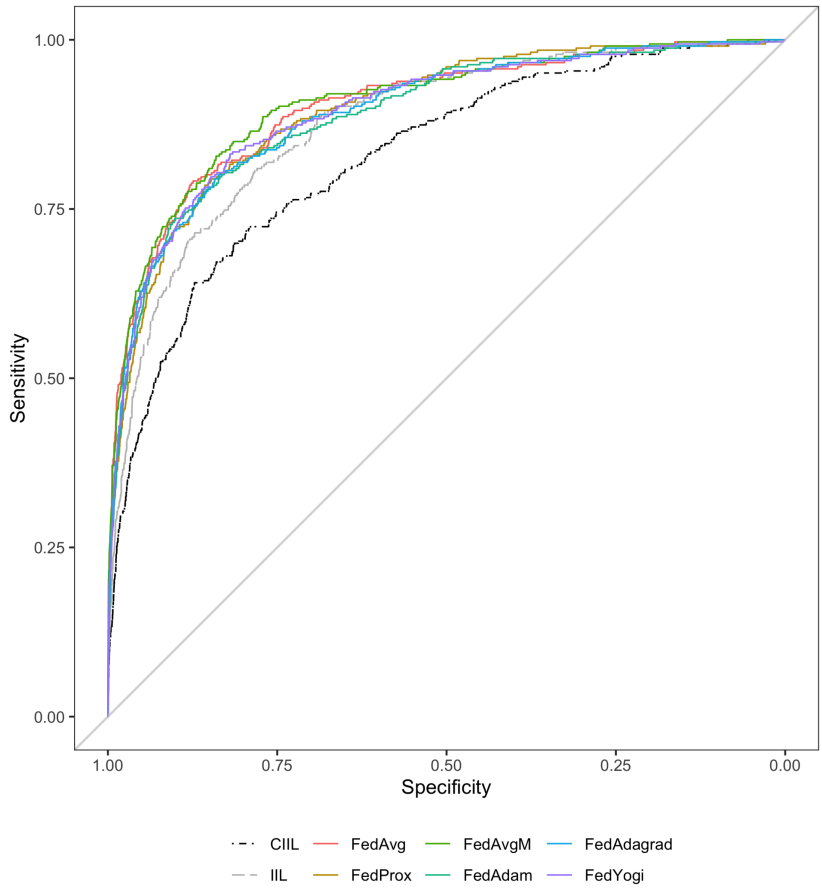

Fig. 3. Comparison of ROC curves obtained with IIL, CIIL, and FL methods for in-hospital mortality prediction.

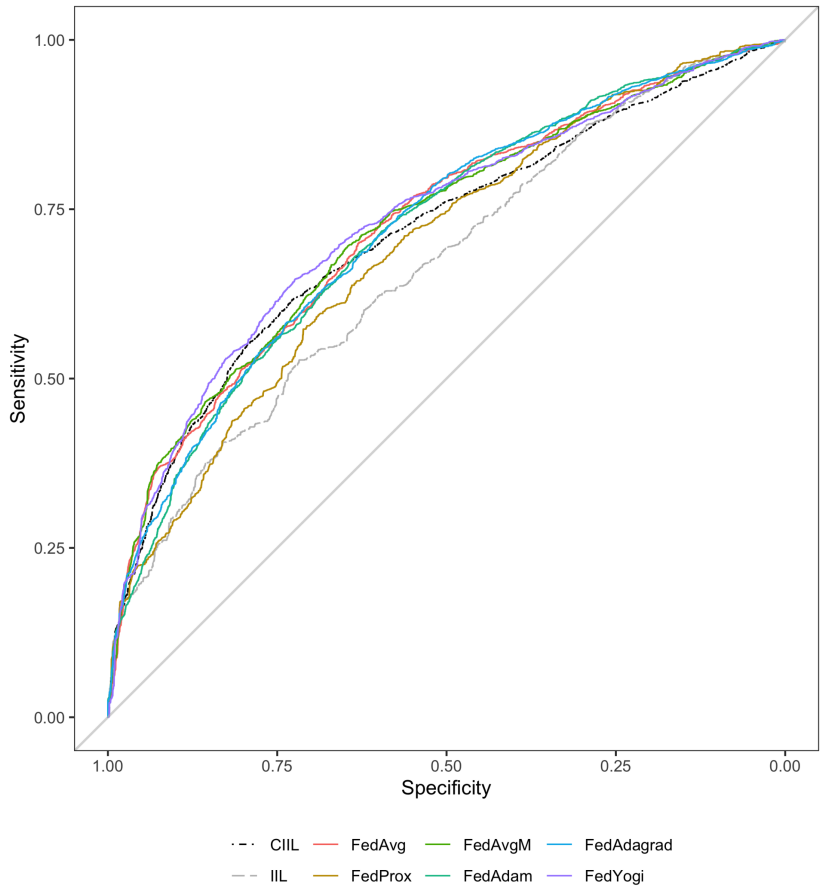

Fig. 4. Comparison of ROC curves obtained with IIL, CIIL, and FL methods for AKI prediction.

REFERENCES

[1] Abbas Acar, Hidayet Aksu, A. Selcuk Uluagac, and Mauro Conti. 2018. A survey on homomorphic encryption schemes: Theory and implementation. *ACM Comput. Surveys* 51, 4 (2018), 1–35.

[2] Erick A. Perez Alday, Annie Gu, Amit J. Shah, Chad Robichaux, An-Kwok Ian Wong, Chengyu Liu, Feifei Liu, Ali Bahrami Rad, Andoni Elola, Salman Seyedi, et al. 2020. Classification of 12-lead ecgs: The physionet/computing in cardiology challenge 2020. *Physiol. Measure.* 41, 12 (2020), 124003.

[3] George J. Annas. 2003. HIPAA regulations—A new era of medical-record privacy? *N. Engl. J. Med.* 348, 15 (2003), 1486–1490.

[4] Yoshinori Aono, Takuya Hayashi, Le Trieu Phong, and Lihua Wang. 2016. Scalable and secure logistic regression via homomorphic encryption. In *Proceedings of the 6th ACM Conference on Data and Application Security and Privacy*. 142–144.

[5] Yoshinori Aono, Takuya Hayashi, Lihua Wang, Shiho Moriai, et al. 2017. Privacy-preserving deep learning via additively homomorphic encryption. *IEEE Trans. Info. Forens. Secur.* 13, 5 (2017), 1333–1345.

[6] Aya Awad, Mohamed Bader-El-Den, James McNicholas, and Jim Briggs. 2017. Early hospital mortality prediction of intensive care unit patients using an ensemble learning approach. *Int. J. Med. Inform.* 108 (2017), 185–195.

[7] Rinaldo Bellomo, Claudio Ronco, John A. Kellum, Ravindra L. Mehta, and Paul Palevsky. 2004. Acute renal failure–definition, outcome measures, animal models, fluid therapy and information technology needs: The second international consensus conference of the acute dialysis quality initiative (ADQI) group. *Crit. Care* 8, 4 (2004), 1–9.

[8] Sabri Boughorbel, Fethi Jarray, Neethu Venugopal, Shabir Moosa, Haithum Elhadi, and Michel Makhlouf. 2019. Federated uncertainty-aware learning for distributed hospital ehr data. Retrieved from https://arXiv:1910.12191.

[9] Kendrick Boyd, Kevin H. Eng, and C. David Page. 2013. Area under the precision-recall curve: Point estimates and confidence intervals. In *Proceedings of the Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 451–466.
[10] Theodora S. Brisimi, Ruidi Chen, Theofanie Mela, Alex Olshevsky, Ioannis Ch Paschalidis, and Wei Shi. 2018. Federated learning of predictive models from federated electronic health records. *Int. J. Med. Inform.* 112 (2018), 59–67.
[11] Ken Chang, Niranjan Balachandar, Carson Lam, Darvin Yi, James Brown, Andrew Beers, Bruce Rosen, Daniel L. Rubin, and Jayashree Kalpathy-Cramer. 2018. Distributed deep learning networks among institutions for medical imaging. *J. Amer. Med. Inform. Assoc.* 25, 8 (2018), 945–954.
[12] Min Chen, Yongfeng Qian, Jing Chen, Kai Hwang, Shiwen Mao, and Long Hu. 2020. Privacy protection and intrusion avoidance for cloudlet-based medical data sharing. *IEEE Trans. Cloud Comput.* 8, 4 (2020), 1274–1283.
[13] Glenn M. Chertow, Elisabeth Burdick, Melissa Honour, Joseph V. Bonventre, and David W. Bates. 2005. Acute kidney injury, mortality, length of stay, and costs in hospitalized patients. *J. Amer. Soc. Nephrol.* 16, 11 (2005), 3365–3370.
[14] Trung Kien Dang, Kwan Chet Tan, Mark Choo, Nicholas Lim, Jianshu Weng, and Mengling Feng. 2020. Building ICU In-hospital mortality prediction model with federated learning. In *Federated Learning*. Springer, 255–268.
[15] Elizabeth R. DeLong, David M. DeLong, and Daniel L. Clarke-Pearson. 1988. Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. *Biometrics* (1988), 837–845.
[16] Hao Du, Ziyuan Pan, Kee Yuan Ngiam, Fei Wang, Ping Shum, and Mengling Feng. 2021. Self-correcting recurrent neural network for acute kidney injury prediction in critical care. *Health Data Sci.* 2021, Article 9808426 (2021), 10 pages.
[17] Rui Duan, Mary Regina Boland, Zixuan Liu, Yue Liu, Howard H. Chang, Hua Xu, Haitao Chu, Christopher H. Schmid, Christopher B. Forrest, John H. Holmes, et al. 2020. Learning from electronic health records across multiple sites: A communication-efficient and privacy-preserving distributed algorithm. *J. Amer. Med. Inform. Assoc.* 27, 3 (2020), 376–385.
[18] Cynthia Dwork. 2011. A firm foundation for private data analysis. *Commun. ACM* 54, 1 (2011), 86–95.
[19] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. 2006. Calibrating noise to sensitivity in private data analysis. In *Proceedings of the Theory of Cryptography Conference*. Springer, 265–284.
[20] Cynthia Dwork, Aaron Roth, et al. 2014. The algorithmic foundations of differential privacy. *Found. Trends Theor. Comput. Sci.* 9, 3–4 (2014), 211–407.
[21] Khaled El Emam, Elizabeth Jonker, Luk Arbuckle, and Bradley Malin. 2011. A systematic review of re-identification attacks on health data. *PLoS One* 6, 12 (2011), e28071.
[22] Rachael L. Fleurence, Lesley H. Curtis, Robert M. Califf, Richard Platt, Joe V. Selby, and Jeffrey S. Brown. 2014. Launching PCORnet, a national patient-centered clinical research network. *J. Amer. Med. Inform. Assoc.* 21, 4 (2014), 578–582.
[23] Robert M. French. 1999. Catastrophic forgetting in connectionist networks. *Trends Cogn. Sci.* 3, 4 (1999), 128–135.
[24] Jerome Friedman, Trevor Hastie, Robert Tibshirani, et al. 2001. *The Elements of Statistical Learning*. Springer Series in Statistics, Vol. 1. Springer, New York.
[25] Joseph Futoma, Morgan Simons, Trishan Panch, Finale Doshi-Velez, and Leo Anthony Celi. 2020. The myth of generalisability in clinical research and machine learning in health care. *Lancet Dig. Health* 2, 9 (2020), e489–e492.
[26] Joana Gameiro, Tiago Branco, and José António Lopes. 2020. Artificial intelligence in acute kidney injury risk prediction. *J. Clin. Med.* 9, 3 (2020), 678.
[27] Lawrence O. Gostin. 2001. National health information privacy: Regulations under the health insurance portability and accountability act. *JAMA* 285, 23 (2001), 3015–3021.
[28] Matei Grama, Maria Musat, Luis Muñoz-González, Jonathan Passerat-Palmbach, Daniel Rueckert, and Amir Alansary. 2020. Robust aggregation for adaptive privacy preserving federated learning in healthcare. Retrieved from https://arXiv:2009.08294.
[29] Otkrist Gupta and Ramesh Raskar. 2018. Distributed learning of deep neural network over multiple agents. *J. Netw. Comput. Appl.* 116 (2018), 1–8.
[30] Tzu-Ming Harry Hsu, Hang Qi, and Matthew Brown. 2019. Measuring the effects of non-identical data distribution for federated visual classification. Retrieved from https://arXiv:1909.06335.
[31] Tzu-Ming Harry Hsu, Hang Qi, and Matthew Brown. 2020. Federated visual classification with real-world data distribution. Retrieved from https://arXiv:2003.08082.
[32] Li Huang, Andrew L. Shea, Huining Qian, Aditya Masurkar, Hao Deng, and Dianbo Liu. 2019. Patient clustering improves efficiency of federated machine learning to predict mortality and hospital stay time using distributed electronic medical records. *J. Biomed. Inform.* 99 (2019), 103291.
[33] Li Huang, Yifeng Yin, Zeng Fu, Shifa Zhang, Hao Deng, and Dianbo Liu. 2020. LoAdaBoost: Loss-based adaboost federated machine learning with reduced computational complexity on IID and non-IID intensive care data. *PLoS One* 15, 4 (2020), e0230706.
[34] Stephanie L. Hyland, Martin Faltys, Matthias Hüser, Xinrui Lyu, Thomas Gumbsch, Cristóbal Esteban, Christian Bock, Max Horn, Michael Moor, Bastian Rieck, et al. 2020. Early prediction of circulatory failure in the intensive care unit using machine learning. *Nature Med.* 26, 3 (2020), 364–373.

[35] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. 2019. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 590–597.

[36] Alistair E. W. Johnson, Tom J. Pollard, Seth J. Berkowitz, Nathaniel R. Greenbaum, Matthew P. Lungren, Chih-ying Deng, Roger G. Mark, and Steven Horng. 2019. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Sci. Data* 6, 1 (2019), 1–8.

[37] Alistair E. W. Johnson, Tom J. Pollard, and Tristan Naumann. 2018. Generalizability of predictive models for intensive care unit patients. In *Proceedings of the Machine Learning for Health (ML4H) Workshop (NeurIPS'18)*. Retrieved from http://arxiv.org/abs/1812.02275.

[38] Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank Reddi, Sebastian Stich, and Ananda Theertha Suresh. 2020. SCAFFOLD: Stochastic controlled averaging for federated learning. In *Proceedings of the International Conference on Machine Learning*. PMLR, 5132–5143.

[39] Rohit J. Kate, Ruth M. Perez, Debesh Mazumdar, Kalyan S. Pasupathy, and Vani Nilakantan. 2016. Prediction and detection models for acute kidney injury in hospitalized older adults. *BMC Med. Inform. Decis. Mak.* 16, 1 (2016), 1–11.

[40] Miran Kim, Yongsoo Song, Shuang Wang, Yuhou Xia, and Xiaoqian Jiang. 2018. Secure logistic regression based on homomorphic encryption: Design and evaluation. *JMIR Med. Inform.* 6, 2 (2018), e19.

[41] Yejin Kim, Jimeng Sun, Hwanjo Yu, and Xiaoqian Jiang. 2017. Federated tensor factorization for computational phenotyping. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 887–895.

[42] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. 2017. Overcoming catastrophic forgetting in neural networks. *Proc. Natl. Acad. Sci. U.S.A.* 114, 13 (2017), 3521–3526.

[43] Jakub Konečnỳ, H. Brendan McMahan, Daniel Ramage, and Peter Richtárik. 2016. Federated optimization: Distributed machine learning for on-device intelligence. Retrieved from https://arXiv:1610.02527.

[44] Junghye Lee, Jimeng Sun, Fei Wang, Shuang Wang, Chi-Hyuck Jun, and Xiaoqian Jiang. 2018. Privacy-preserving patient similarity learning in a federated environment: Development and analysis. *JMIR Med. Inform.* 6, 2 (2018), e20.

[45] Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. 2018. Federated optimization in heterogeneous networks. Retrieved from https://arXiv:1812.06127.

[46] Craig M. Lilly, John M. McLaughlin, Huifang Zhao, Stephen P. Baker, Shawn Cody, Richard S. Irwin, and UMass Memorial Critical Care Operations Group. 2014. A multicenter study of ICU telemedicine reengineering of adult critical care. *Chest* 145, 3 (2014), 500–507.

[47] Dianbo Liu, Dmitriy Dligach, and Timothy Miller. 2019. Two-stage federated phenotyping and patient representation learning. Retrieved from https://arXiv:1908.05596.

[48] Songtao Lu, Yawen Zhang, and Yunlong Wang. 2020. Decentralized federated learning for electronic health records. In *Proceedings of the 54th Annual Conference on Information Sciences and Systems (CISS'20)*. IEEE, 1–5.

[49] Clement J. McDonald, J. Marc Overhage, Michael Barnes, Gunther Schadow, Lonnie Blevins, Paul R. Dexter, Burke Mamlin, and INPC Management Committee. 2005. The indiana network for patient care: A working local health information infrastructure. *Health Affairs* 24, 5 (2005), 1214–1220.

[50] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. 2017. Communication-efficient learning of deep networks from decentralized data. In *Artificial Intelligence and Statistics*. PMLR, 1273–1282.

[51] Silvio Micali, Oded Goldreich, and Avi Wigderson. 1987. How to play any mental game. In *Proceedings of the 19th ACM Symposium on Theory of Computing (STOC'87)*. ACM, 218–229.

[52] Riccardo Miotto, Fei Wang, Shuang Wang, Xiaoqian Jiang, and Joel T. Dudley. 2018. Deep learning for healthcare: Review, opportunities, and challenges. *Brief.Bioinform.* 19, 6 (2018), 1236–1246.

[53] Hamid Mohamadlou, Anna Lynn-Palevsky, Christopher Barton, Uli Chettipally, Lisa Shieh, Jacob Calvert, Nicholas R. Saber, and Ritankar Das. 2018. Prediction of acute kidney injury with a machine learning algorithm using electronic health record data. *Can. J. Kidney Health Dis.* 5 (2018), 2054358118776326.

[54] Luis Muñoz-González, Kenneth T. Co, and Emil C. Lupu. 2019. Byzantine-robust federated machine learning through adaptive model averaging. Retrieved from https://arXiv:1909.05125.

[55] Shawn N. Murphy, Griffin Weber, Michael Mendis, Vivian Gainer, Henry C. Chueh, Susanne Churchill, and Isaac Kohane. 2010. Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). *J. Amer. Med. Inform. Assoc.* 17, 2 (2010), 124–130.

[56] Dianwen Ng, Xiang Lan, Melissa Min-Szu Yao, Wing P. Chan, and Mengling Feng. 2021. Federated learning: A collaborative effort to achieve better medical imaging models for individual sites that have small labelled datasets. *Quant. Imag. Med. Surg.* 11, 2 (2021), 852.

[57] Valeria Nikolaenko, Udi Weinsberg, Stratis Ioannidis, Marc Joye, Dan Boneh, and Nina Taft. 2013. Privacy-preserving ridge regression on hundreds of millions of records. In *Proceedings of the IEEE Symposium on Security and Privacy*. IEEE, 334–348.

[58] Joshua Parreco, Hahn Soe-Lin, Jonathan J. Parks, Saskya Byerly, Matthew Chatoor, Jessica L. Buicko, Nicholas Namias, and Rishi Rattan. 2019. Comparing machine learning algorithms for predicting acute kidney injury. *Amer. Surg.* 85, 7 (2019), 725–729.

[59] Stephen R. Pfohl, Andrew M. Dai, and Katherine Heller. 2019. Federated and differentially private learning for electronic health records. Retrieved from https://arXiv:1911.05861.

[60] Tom J. Pollard, Alistair E. W. Johnson, Jesse D. Raffa, Leo A. Celi, Roger G. Mark, and Omar Badawi. 2018. The eICU collaborative research database, a freely available multi-center database for critical care research. *Sci. Data* 5, 1 (2018), 1–13.

[61] Sanjay Purushotham, Chuizheng Meng, Zhengping Che, and Yan Liu. 2018. Benchmarking deep learning models on large healthcare datasets. *J. Biomed. Inform.* 83 (2018), 112–134.

[62] Sashank Reddi, Zachary Charles, Manzil Zaheer, Zachary Garrett, Keith Rush, Jakub Konečnỳ, Sanjiv Kumar, and H. Brendan McMahan. 2020. Adaptive federated optimization. Retrieved from https://arXiv:2003.00295.

[63] Richard D. Riley, Joie Ensor, Kym I. E. Snell, Thomas P. A. Debray, Doug G. Altman, Karel G. M. Moons, and Gary S. Collins. 2016. External validation of clinical prediction models using big datasets from e-health records or IPD meta-analysis: Opportunities and challenges. *bmj* 353 (2016).

[64] Ronald L. Rivest, Len Adleman, Michael L. Dertouzos, et al. 1978. On data banks and privacy homomorphisms. *Found. Secure Comput.* 4, 11 (1978), 169–180.

[65] Selva Muthu Kumaran Sathappan, Young Seok Jeon, Trung Kien Dang, Su Chi Lim, Yi-Ming Shao, E. Shyong Tai, and Mengling Feng. 2021. Transformation of electronic health records and questionnaire data to OMOP CDM: A feasibility study using SG_T2DM dataset. *Appl. Clin. Inform.* 12, 4 (2021), 757–767.

[66] Pulkit Sharma, Farah E. Shamout, and David A. Clifton. 2019. Preserving patient privacy while training a predictive model of in-hospital mortality. Retrieved from https://arXiv:1912.00354.

[67] Micah J. Sheller, Brandon Edwards, G. Anthony Reina, Jason Martin, Sarthak Pati, Aikaterini Kotrotsou, Mikhail Milchenko, Weilin Xu, Daniel Marcus, Rivka R. Colen, et al. 2020. Federated learning in medicine: Facilitating multi-institutional collaborations without sharing patient data. *Sci. Rep.* 10, 1 (2020), 1–12.

[68] Micah J. Sheller, G. Anthony Reina, Brandon Edwards, Jason Martin, and Spyridon Bakas. 2018. Multi-institutional deep learning modeling without sharing patient data: A feasibility study on brain tumor segmentation. In *Proceedings of the International MICCAI Brainlesion Workshop*. Springer, 92–104.

[69] Daniel Sonntag, Volker Tresp, Sonja Zillner, Alexander Cavallaro, Matthias Hammon, André Reis, Peter A. Fasching, Martin Sedlmayr, Thomas Ganslandt, Hans-Ulrich Prokosch, et al. 2016. The clinical data intelligence project. *Informatik-Spektrum* 39, 4 (2016), 290–300.

[70] Peter J. Stow, Graeme K. Hart, Tracey Higlett, Carol George, Robert Herkes, David McWilliam, Rinaldo Bellomo, ANZICS Database Management Committee, et al. 2006. Development and implementation of a high-quality clinical database: the australian and new zealand intensive care society adult patient database. *J. Crit. Care* 21, 2 (2006), 133–141.

[71] Scott M. Sutherland, Lakhmir S. Chawla, Sandra L. Kane-Gill, Raymond K. Hsu, Andrew A. Kramer, Stuart L. Goldstein, John A. Kellum, Claudio Ronco, Sean M. Bagshaw, and 15 ADQI Consensus Group. 2016. Utilizing electronic health records to predict acute kidney injury risk and outcomes: Workgroup statements from the 15th ADQI consensus conference. *Can. J. Kidney Health Dis.* 3 (2016), 99.

[72] Xiaoqing Tan, Chung-Chou H. Chang, and Lu Tang. 2021. A tree-based federated learning approach for personalized treatment effect estimation from heterogeneous data sources. Retrieved from https://arXiv:2103.06261.

[73] Chandra Thapa and Seyit Camtepe. 2021. Precision health data: Requirements, challenges and existing techniques for data security and privacy. *Comput. Biol. Med.* 129 (2021), 104130.

[74] Patrick J. Thoral, Jan M. Peppink, Ronald H. Driessen, Eric J. G. Sijbrands, Erwin J. O. Kompanje, Lewis Kaplan, Heatherlee Bailey, Jozef Kesecioglu, Maurizio Cecconi, Matthew Churpek, et al. 2021. Sharing ICU patient data responsibly under the Society of Critical Care Medicine/European Society of Intensive Care Medicine joint data science collaboration: The Amsterdam University Medical Centers Database (AmsterdamUMCdb) example. *Crit. Care Med.* 49, 6 (2021), e563.

[75] Volker Tresp, J. Marc Overhage, Markus Bundschus, Shahrooz Rabizadeh, Peter A. Fasching, and Shipeng Yu. 2016. Going digital: A survey on digitalization and large-scale data analytics in healthcare. *Proc. IEEE* 104, 11 (2016), 2180–2206.

[76] Anup Tuladhar, Sascha Gill, Zahinoor Ismail, Nils D. Forkert, Alzheimer's Disease Neuroimaging Initiative, et al. 2020. Building machine learning models without sharing patient data: A simulation-based analysis of distributed learning by ensembling. *J. Biomed. Inform.* 106 (2020), 103424.

[77] Akhil Vaid, Suraj K. Jaladanki, Jie Xu, Shelly Teng, Arvind Kumar, Samuel Lee, Sulaiman Somani, Ishan Paranjpe, Jessica K. De Freitas, Tingyi Wanyan, 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 (2021), e24207.

[78] Praneeth Vepakomma, Otkrist Gupta, Tristan Swedish, and Ramesh Raskar. 2018. Split learning for health: Distributed deep learning without sharing raw patient data. Retrieved from https://arXiv:1812.00564.
[79] Paul Voigt and Axel Von dem Bussche. 2017. The EU general data protection regulation (GDPR). *A Practical Guide*, vol. 10, 1st ed. Springer, Cham, 3152676.
[80] Erica A. Voss, Rupa Makadia, Amy Matcho, Qianli Ma, Chris Knoll, Martijn Schuemie, Frank J. DeFalco, Ajit Londhe, Vivienne Zhu, and Patrick B. Ryan. 2015. Feasibility and utility of applications of the common data model to multiple, disparate observational health databases. *J. Amer. Med. Inform. Assoc.* 22, 3 (2015), 553–564.
[81] L. Wynants, D. M. Kent, D. Timmerman, C. M. Lundquist, and B. Van Calster. 2019. Untapped potential of multicenter studies: A review of cardiovascular risk prediction models revealed inappropriate analyses and wide variation in reporting. *Diagnost. Prognost. Res.* 3, 1 (2019), 1–17.
[82] Cao Xiao, Edward Choi, and Jimeng Sun. 2018. Opportunities and challenges in developing deep learning models using electronic health records data: A systematic review. *J. Amer. Med. Inform. Assoc.* 25, 10 (2018), 1419–1428.
[83] Jie Xu, Benjamin S. Glicksberg, Chang Su, Peter Walker, Jiang Bian, and Fei Wang. 2021. Federated learning for healthcare informatics. *J. Healthcare Inform. Res.* 5, 1 (2021), 1–19.
[84] Jie Xu, Zhenxing Xu, Peter Walker, and Fei Wang. 2020. Federated patient hashing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 6486–6493.
[85] Zeyue Xue, Pan Zhou, Zichuan Xu, Xiumin Wang, Yulai Xie, Xiaofeng Ding, and Shiping Wen. 2021. A resource-constrained and privacy-preserving edge-computing-enabled clinical decision system: A federated reinforcement learning approach. *IEEE Internet Things J.* 8, 11 (2021), 9122–9138.
[86] Andrew C. Yao. 1982. Protocols for secure computations. In *Proceedings of the 23rd Annual Symposium on Foundations of Computer Science (SFCS'82)*. IEEE, 160–164.

Received May 2021; revised January 2022; accepted January 2022