5-11-2023

Enabling Privacy-Preserving Prediction for Length of Stay in ICU - A Multimodal Federated-Learning-based Approach

Tongnian Wang
The University of Texas at San Antonio, tongnian.wang@utsa.edu

Yuanxiong Guo
The University of Texas at San Antonio, yuanxiong.guo@utsa.edu

Kim-Kwang Raymond Choo
University of Texas at San Antonio, raymond.choo@utsa.edu

ENABLING PRIVACY-PRESERVING PREDICTION FOR LENGTH OF STAY IN ICU - A MULTIMODAL FEDERATED-LEARNING-BASED APPROACH

Complete Research

Tongnian Wang, The University of Texas at San Antonio, USA, tongnian.wang@utsa.edu

Yuanxiong Guo, The University of Texas at San Antonio, USA, yuanxiong.guo@utsa.edu

Kim-Kwang Raymond Choo, The University of Texas at San Antonio, USA, raymond.choo@utsa.edu

Abstract

While the proliferation of data-driven machine learning approaches has resulted in new opportunities for precision healthcare, there are a number of challenges associated with fully utilizing medical data, for example partly due to the heterogeneity of data modalities in electronic health records. Moreover, medical data often sits in data silos due to various regulatory, privacy, ethical, and legal considerations, which complicates efforts to fully utilize machine learning. Motivated by these challenges, we focus on clinical care—length of stay prediction and propose a Multimodal Federated Learning approach. The latter is designed to leverage both privacy-preserving federated learning and multimodal data to facilitate length of stay prediction. By applying this approach to a real-world medical dataset, we demonstrate the predictive power of our approach as well as how it can address the earlier discussed challenges. The findings also suggest the potential of the proposed multimodal federated learning approach for other similar healthcare settings.

Keywords: Length of Stay Prediction, Privacy, Federated Learning, Multimodality.

1 Introduction

Ongoing digitization efforts within the healthcare sector, along with the proliferation of artificial intelligence (AI) technologies, have the potential to transform clinical practice and improve quality of healthcare and medical services, ranging from diagnosis (Gunčar et al., 2018; Myszczynska et al., 2020) to outcome prediction (Awad et al., 2017; Davoodi and Moradi, 2018), and so on. Typical clinical practices often rely on a variety of information formats contained within the electronic health records (EHRs); therefore, utilizing knowledge from various input modalities or data sources in AI-based medical systems to support more accurate and effective healthcare systems (Soenksen et al., 2022). More specifically, it has been shown that leveraging multiple modalities from EHRs such as tabular data (e.g., demographics, admission/discharge details), time-series data (e.g., blood chemistry, respiratory), unstructured sequence data (e.g., notes, written reports), and image data (e.g., x-rays, CT scans), can potentially improve the performance of AI systems in healthcare settings, compared to single-modality for the same task (S.-C. Huang et al., 2020; Soenksen et al., 2022). Consequently, developing multimodal AI-based approaches for medical care is a topic of ongoing interest, despite new and existing challenges (e.g., how to effectively choose the appropriate data modalities, or collect large volumes of patient health data from multiple modalities).

There are additional challenges, for example in terms of achieving an optimal or desired trade-off between improved quality of care and minimizing unintended consequences (e.g., loss of privacy)

(Anderson and Agarwal, 2011). More specifically, data-driven machine learning (ML) approaches are highly reliant on real-world data of different modalities and from different sources. The healthcare sector, however, has exacting restrictions on protecting the privacy of personal health information (PHI) including in precision health applications (Anderson and Agarwal, 2011; B. Liu, Pavlou, and Cheng, 2022; Xu et al., 2021). Such restrictions have been encoded in privacy regulations such as the General Data Protection Regulation (GDPR) (Voigt and Von dem Bussche, 2017) and the Health Insurance Portability and Accountability Act (HIPAA) (Department of Health and Human Services, 2013), which can limit the sharing of data between organizations. In addition, there are other operational and organizational considerations. For example, collecting, curating, and maintaining a high-quality medical dataset take considerable time, effort, and costs (Rieke et al., 2020). Consequently, to achieve clinical-grade precision for ML-based clinical applications, it can be operationally challenging to compile large-scale curated datasets without requiring the data collector to hand over fine-grained control over such datasets (De Fauw et al., 2018; Rieke et al., 2020; F. Wang, Casalino, and Khullar, 2019).

One viable solution is the local-only approach, where each hospital or a group of several hospitals in a small region only trains a ML-based health system solely on their purposefully curated dataset(s) (Panch, Mattie, and Atun, 2019; Rieke et al., 2020). One disadvantage of such an approach is that healthcare organizations or other stakeholder groups cannot extract and leverage knowledge and intelligence beyond their organization (Y.-K. Lin, M. Lin, and H. Chen, 2019). This can introduce biases to ML-based algorithms, and utilizing such algorithms can exacerbate existing health disparities (Panch, Mattie, and Atun, 2019; Panch, Mattie, and Celi, 2019). In other words, a poorly designed AI-based health system may result in discrimination among various socio-demographic groups, leading to unequal outcomes (Ghassemi et al., 2020; Leslie et al., 2021; Obermeyer et al., 2019). These challenges reinforce the importance of designing large-scale AI-assisted healthcare systems while adhering to privacy regulations.

Recent efforts have been made to develop federated learning (FL)-based methods to address the above-discussed challenges. More specifically, FL enables the ML process to occur locally at each participating institution without moving patient data beyond the firewalls of the institutions in which they reside. Instead, only model characteristics such as parameters are exchanged to achieve the learning objective. Participants collaboratively train models under the coordination of a central server, and eventually, an optimized consensus model can be obtained. FL implicitly offers a certain degree of privacy in all cases because participants never directly access data from others and only receive model parameters that are aggregated over multiple participants (McMahan et al., 2017). Although FL was originally designed for privacy protection, it also empowers data controllers to establish their data governance processes, control data access, and revoke it, which has striking implications for healthcare (Rieke et al., 2020). Accordingly, establishing FL on a global scale could ensure high-quality clinical support regardless of the treatment location, and such large-scale collaborative learning can reduce the sampling biases in less representative datasets (Rieke et al., 2020).

In this study, we focus on an important problem in clinical care, which is the remaining length of stay (LOS) prediction in intensive care units (ICU). The accurate prediction of LOS in ICUs can facilitate bed management and cost control, and improve patient outcomes. Not surprisingly, LOS is a key parameter for patients, clinicians, and hospital administrators (Bacchi et al., 2022; Bartel, Chan, and Kim, 2020). For example, Rotter et al. (2010) have shown that the longer a patient stays in the hospital, the less likely (s)he would have a positive outcome. When a patient is kept in a bed longer than necessary, the bed is not available for other patients (Bacchi et al., 2022; Daghistani et al., 2019). However, predicting LOS can be influenced by many factors, particularly for patients with complex medical conditions / history (Bacchi et al., 2022). We also observe that existing LOS prediction approaches have varying limitations. For example, many of these approaches were designed to support one prediction throughout the entire ICU stay, and did not make full use of the heterogeneous data modalities in EHRs. Moreover, there is a lack of fine-grained approaches to address privacy and data governance problems in real-world scenarios. These observations lead us to explore the following research question (RQ):

RQ: How can we enhance LOS prediction for ICU using multimodal data from different data sources (e.g., different hospitals) while also safeguarding patient information privacy?

Our work contributes to the body of knowledge in two ways. First, motivated by properly addressing the problems associated with the implementation of centralized ML in real-world applications, we use the design science research (DSR) paradigm (Hevner et al., 2004; Peffers et al., 2007) to guide the design, development, and evaluation of our proposed multimodal federated-learning-based method (i.e., the artifact), with the objective of supporting large-scale clinical prediction training and, instantiates in LOS prediction while considering data governance and privacy issues. Second, by implementing and evaluating our method on a real-world dataset, we provide insights into the feasibility and practicability of the proposed artifact. Researchers and practitioners can use our proposed approach in real-world applications and generalize it to other clinical tasks as well.

Our paper is structured as follows. In the second section, we briefly review the extant literature on LOS prediction and approaches that also support data governance and privacy. In the third section, we describe our research methodology, which follows the design science paradigm. The workflow and components of our system and the data preparation process are presented in the fourth section. We evaluate our final artifact on a real-world dataset in the fifth section, prior to discussing the theoretical and practical implications, as well as limitations and future research possibilities, in the sixth section. We summarize this work in the final section.

2 Related Work

2.1 Length of Stay Prediction

Patients' LOS in ICUs is generally considered a key metric for ICU resource utilization and patient outcome evaluation (Bacchi et al., 2022; Bartel, Chan, and Kim, 2020). However, LOS may be influenced by many factors, particularly in complex medical patients, and may be difficult to predict (Bacchi et al., 2022). In most existing studies, patients were usually divided into two or multiple groups based on their LOS, aiming at identifying the risk for long stays (Alsinglawi et al., 2022; Daghistani et al., 2019; Hachesu et al., 2013). This kind of binary or multi-class classification problems ignored that LOS is naturally formulated as a regression task. Some existing literature studied the more challenging regression task but their methods only conducted one prediction throughout a patient's entire ICU stay, which did not make full use of the time-series data (Baek et al., 2018; Sotoodeh and Ho, 2019; Tsai et al., 2016). Therefore, considering the remaining LOS as a regression at each time step would be more important for efficient scheduling and ICU resource management (Harutyunyan et al., 2019).

Besides, most previous literature has leveraged traditional ML-based models to perform LOS prediction, such as logistic regression (Alsinglawi et al., 2022; Tsai et al., 2016), support vector machine (Daghistani et al., 2019; Hachesu et al., 2013), random forest (Alsinglawi et al., 2022; Baek et al., 2018; Morton et al., 2014), or simple deep neural networks (DNNs) such as multi-layer perceptrons (Daghistani et al., 2019; Hachesu et al., 2013; Tsai et al., 2016). However, the performance of such approaches varied a lot, and these papers did not consider time-series features, which are almost the most relevant features for tasks like LOS prediction (Soenksen et al., 2022). Later, DNNs such as LSTM-based models (Harutyunyan et al., 2019; Rajkomar et al., 2018), attention-based models (Song et al., 2018), and even temporal convolution-based models (Rocheteau, Liò, and Hyland, 2021), which are capable of handling dependence across the time domain are proven to perform better on LOS prediction than other models. However, these studies focused more on how to design models that can leverage time-series features for LOS prediction, while the more important problem is how to leverage time-series data and tabular data for multimodal LOS prediction, because features like gender, age, ethnicity, etc., contain deep-seated patterns of health discrimination, which play an important role in identifying health inequities caused by biased AI systems.

2.2 Privacy and Clinical Prediction

Existing research on AI and health information privacy has indeed emphasized that the complexity of the medical context in terms of the plurality of stakeholders (Beckerman et al., 2008), coupled with the highly personal and sensitive nature of personal health information (Trumbo, McComas, and Kannaovakun, 2007) suggest that investigations of privacy must pay attention to a broad range of risk elements. To address these challenges, several privacy mechanisms have been proposed to manage privacy issues in AI. Specifically, de-identification (or anonymization) (Sedayao, Bhardwaj, and Gorade, 2014; Sweeney, 2000) has been widely used, as it can help to mitigate privacy issues by removing explicit identifiers such as name and address, or by replacing them with pseudonyms or codes that cannot be traced back to an individual. However, it is important to note that achieving complete de-identification can be challenging, and even if a dataset is "anonymized", patient privacy may still be put at risk as even seemingly insignificant information could potentially be used for the re-identification of patients (Abouelmehdi, Beni-Hessane, and Khaloufi, 2018; McMahan et al., 2017; Sweeney, 2000). In addition, differential privacy (Dwork et al., 2006), is another privacy-preserving technique to safeguard sensitive information by introducing random noise to the data to avoid the identification of individuals. Nevertheless, this technique may lower the accuracy of ML models and may not be effective against all types of attacks. For example, if an attacker has access to additional external information about an individual in the dataset, they may still be able to identify that individual despite the random noise added by differential privacy. Cryptographic approaches (Clifton et al., 2002; Laur, Lipmaa, and Mielikäinen, 2006) often bring in huge extra computation overheads and costs with just a certain level of provable privacy (Abouelmehdi, Beni-Hessane, and Khaloufi, 2018; Duan, Canny, and Zhan, 2010).

On the other hand, federated learning (FL) (McMahan et al., 2017), a learning paradigm where multiple entities collaborate in solving a ML problem without exchanging their original data, would be a potential solution to the privacy problem in the healthcare domain. The majority of existing FL research focuses on FL optimization and addressing challenges associated with FL itself, while in medical, there still lacks evidence about their business value in real-world applications. For example, J. Lee et al. (2018) proposed a federated patient hashing framework which can compute the similarities between patients across institutions, and demonstrated the accuracy and usability of the proposed framework. However, the proposed system can not be generalized to other problems, and lacked analysis about the social impacts. Some prior studies also applied FL-based methods to other medical problems, for example, mortality prediction (G. Lee and Shin, 2020; Vaid et al., 2021), and hospitalizations prediction (Brisimi et al., 2018), but with the purposes of preserving privacy and improving accuracy or efficiency. These studies either evaluated their methods on only a few hospitals, or included very limited input features, which might not be able to handle the complexity of the medical context, as we argued before. In the context of LOS prediction, L. Huang et al. (2019) presented a federated machine learning model for mortality and prolonged LOS classification only based on drug features, and Pfohl, Dai, and Heller (2019) studied the efficacy of FL setting on a simple prolonged LOS prediction task based on simple ML models that cannot handle time-series data properly. In this case, there is still lack of studies which leveraged multimodal data by taking time-series data into account, as they are important for LOS prediction, or even other clinical prediction tasks. To conclude the status of LOS prediction research, there is a gap in effective and generalized approaches for accurate LOS prediction systems that can overcome privacy and data governance challenges (privacy-preserving) while simultaneously utilizing multimodal data to improve prediction accuracy (multimodal). Therefore, a multimodal federated-learning-based approach would be a promising approach for LOS prediction.

3 Methodology

We followed the Design Science Research (DSR) approach which is a fundamental paradigm in IS research concerned with the construction of socio-technical artifacts to solve organizational and societal problems and derive prescriptive design knowledge (Gregor and Hevner, 2013; Hevner et al., 2004; Peffers et al., 2007). In line with DSR, we followed the seven DSR guidelines by Hevner et al. (2004) as well as the DSR methodology proposed by Peffers et al. (2007) for developing our IT artifact.

As no research has suggested a multimodal federated-learning-based method for real-time remaining LOS prediction so far, our goal is to develop a method that enables the use of multimodal EHR data to perform privacy-preserving prediction on real-time LOS in the ICUs. According to Gregor and Hevner (2013), the developed artifact within DSR can be considered as knowledge contribution if the artifact either provides a new solution for a known problem, a new solution for new problems, or extent known solutions to new problems. In our case, our proposed multimodal federated LOS prediction method can be considered as a new solution for a known problem of LOS prediction. After the problem identification, we examined the existing approaches for LOS prediction. Subsequently, we began the design and development process in which we developed our method for LOS prediction based on federated learning in a way that allows for leveraging multimodal EHR data for more convincing prediction. Then the artifact was demonstrated and evaluated (Hevner et al., 2004; Peffers et al., 2007). We evaluated our method by applying it to a real-world medical dataset and investigated the efficacy of our method. Finally, for communication, our work offers prescriptive knowledge on how multimodal privacy-preserving learning methods for LOS prediction ought to be designed, and it also has a variety of practical implications for different stakeholders (Hevner et al., 2004; Peffers et al., 2007).

4 Multimodal Federated Length of Stay Prediction

We now discuss in detail our framework, which supports the privacy-preserving prediction of patient accurate remaining LOS in the ICUs using multimodal data. In a nutshell, the system involves a central server, which mainly takes care of parameter aggregation, as well as a number of hospitals, each training a neural network model with the same model architecture. Figure 1 outlines the overall workflow of the system as per the following steps, and this iterative learning process will continue until the model convergence:

1. At each communication round, the server randomly selects a fraction of the participating hospitals for federated training of the prediction model.
2. The server sends the parameters of the Temporal Pointwise Convolution (TPC) prediction model, aggregated at the server, to the hospitals (parameters are initialized at random in the first round).
3. Selected hospitals conduct local update on the TPC model using stochastic gradient descent and the local available data based on the received model parameters.
4. The hospitals send back the updated TPC model parameters to the central server.
5. The server aggregates the local model parameters sent by the selected hospitals using a specific FL algorithm to produce the new global model parameters, and the aggregated new global model will be serving as the initial model for the next communication round.

Formally, our task is to predict the remaining LOS at regular timepoints (every hour) throughout a patient's entire stay in the ICU, up to the discharge time T, using features from two data modalities—tabular data and time-series data. After feature extraction, time-series features will be fed into the TPC model, and the output of the TPC model will be combined with features from tabular data as the inputs to a two-layer pointwise convolution model. By doing this, multiple modalities can be used together to facilitate the model performance. The time-series features of a single ICU stay contain various timepoints (t), and there are two channels initially: feature values ($h_{t,1}$), and their corresponding decay indicators ($h_{t,2}$). The decay indicators can tell how recently the observation $h_{t,1}$ was recorded. We will discuss about the model in detail in 4.1.3. In the rest of this section, we describe the system's various entities and components.

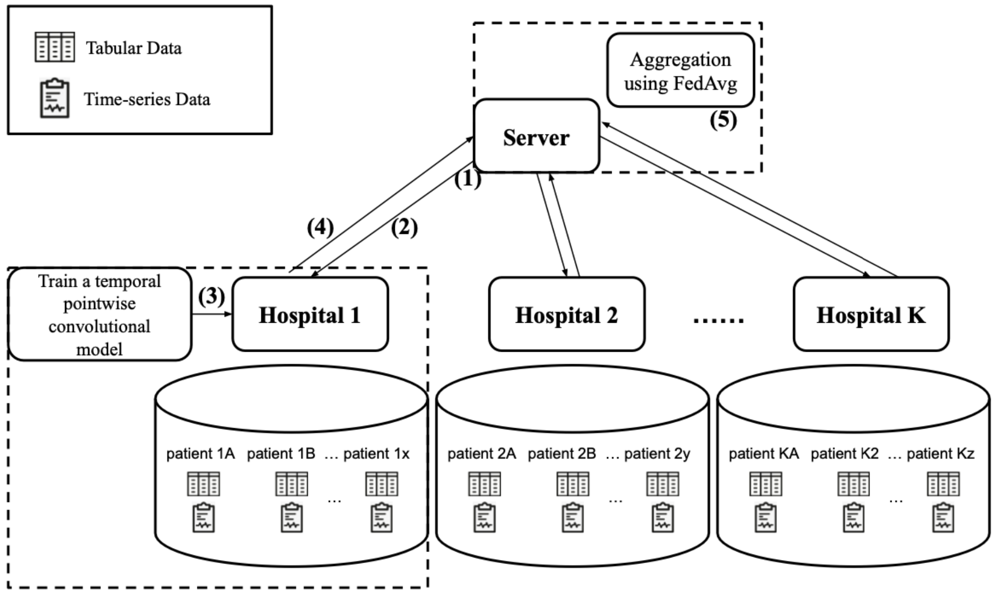

Figure 1. Workflow of the Proposed Federated Multimodal LOS Prediction Framework.

4.1 Components

4.1.1 Hospitals

We operate in a collaborative setting with a number of hospitals engaging to train a federated multimodal TPC model geared to predict remaining LOS in ICUs. Each hospital conducts local training on its locally available data in each communication round, and only model parameters will be sent to the central server.

4.1.2 Server

The central server is responsible for collecting trained local models, aggregating them, and sending the updated global model back to hospitals. The participating hospitals should trust the server for the exchange of model parameters. And in our framework, the aggregation is performed using FedAvg algorithms (McMahan et al., 2017), one of the standard FL algorithms. The FedAvg procedure consists of several steps: the server initializes the model and randomly chooses a subset of hospitals to participate in training. During each communication round, the server sends the model parameters to the selected hospitals. The hospitals perform local updates based on the received model parameters and then send back the updated model parameters to the central server. The server aggregates the local model parameters using weighted averaging to generate the new global model, which serves as the initial model for the next communication round. The above process is repeated for multiple communication rounds until convergence.

4.1.3 Model

As for the model used by each hospital in the local training, we adopted the TPC model proposed by Rocheteau, Liò, and Hyland (2021), which is state-of-the-art ML model for LOS prediction. The overall workflow of the model is shown in Figure 2. The original design of this TPC model focused on how it can handle time-series features in medical data, while we focus on how it can leverage multiple input modalities and deliver more accurate results compared to other ML models.

In general, there are three main fusion strategies to join data from multiple modalities, namely early, joint, and late fusion (S.-C. Huang et al., 2020). Early fusion refers to join multiple input modalities into a single feature vector and feed into a ML model for training, and late fusion is the process of leveraging outputs from multiple trained models to make a final prediction (S.-C. Huang et al., 2020). In our case, joint fusion would be a more suitable approach, because features from tabular data do not really need

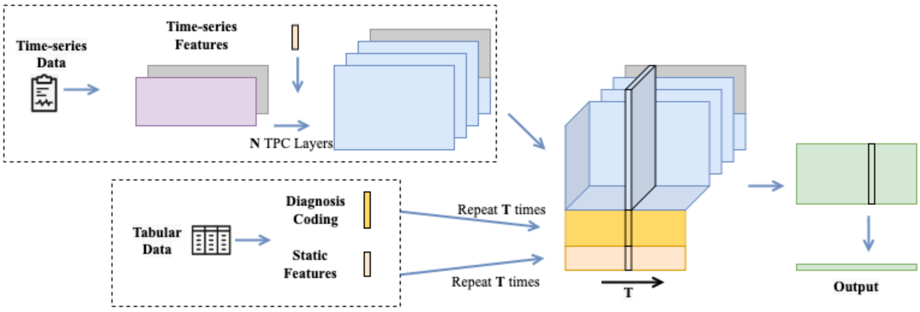

Figure 2. The TPC Model with Multimodal Inputs. Adapted from original paper Rocheteau, Liò, and Hyland (2021). The time-series features $h_{t,1}$ (purple) and corresponding decay indicators $h_{t,2}$ (grey) will be processed by N TPC layers, and then joint fusion is applied to combine them with features from tabular data among the feature domain. Finally a two-layer pointwise convolution model is applied and obtain the final output $\hat{y}$ (green).

any feature extraction models. In other words, feature representations are learned for time-series data first through N TPC layers, and then combined with features from tabular data for final prediction. More specifically, the extracted time-series features $h_{t,1}$ and corresponding decay indicators $h_{t,2}$ are the initial inputs to the first TPC layer, and will be processed by N TPC layers, where the temporal convolution networks (TCN) (Kalchbrenner et al., 2016; Oord et al., 2016) will examine through regular timepoint t and map the X input channels into Y output channels, and the pointwise convolution will be applied separately to each timepoint t with information from static features. Besides, after obtaining the static and diagnosis features from tabular data, they will be combined with time-series representations among the feature domain using joint fusion. Finally, a two-layer pointwise convolution model is implemented, so that final prediction $\hat{y}$ can be obtained.

More specifically, in order to handle time-series features, the TPC model combines temporal convolutional layers (Kalchbrenner et al., 2016), which can capture the causal dependencies across the time domain, as well as pointwise convolutional layers (M. Lin, Q. Chen, and Yan, 2013), which can compute higher-level features from interactions in the feature domain. The temporal convolution operation is defined as

$$(f^{n,i} * h^{n,i})(t) = \sum_{j=1}^{k} f^{n,i}[j] \times h^{n,i}_{t-d(j-1)} \tag{1}$$

where $h^{n,i}_{1:t}$ represents the layer n's temporal inputs, up to timepoint t, and $f^{n,i}$ is the convolutional filter for each feature. And the pointwise convolution operation is applied separately to each timepoint, where in the n_{th} layer, it is defined as

$$(g^n * p^n)(t) = \sum_{i=1}^{P^n} g^n[i] \times p^{n,i}_t \tag{2}$$

where g^n represents the pointwise filter, and p^n_t is the interaction features concatenated by several kinds of features across the feature domain. Therefore, this model could extract both temporal trends and inter-feature relationships so that it is suitable for capturing the patient's clinical state, and has been demonstrated to yield state-of-the-art results on LOS prediction (Rocheteau, Liò, and Hyland, 2021). Since our focus is not on the model, please refer to the original paper Rocheteau, Liò, and Hyland (2021) for more details about how this model deals with the time-series modality.

In addition, Figure 3 shows that both total LOS and remaining LOS have significant positive skewed distributions and the remaining LOS has an extreme skew with mean and median values of 3.99 and 1.96 days respectively. In order to train the model for a regression task that has such an extremely right-skewed distribution, we use Mean Squared Logarithmic Error (MSLE) loss as the loss function, to handle the skewness, see below:

$$L(y,\hat{y}) := \frac{1}{N}\sum_{i=0}^{N}\left(\log\left(y_i+1\right)-\log\left(\hat{y}_i+1\right)\right)^2 \tag{3}$$

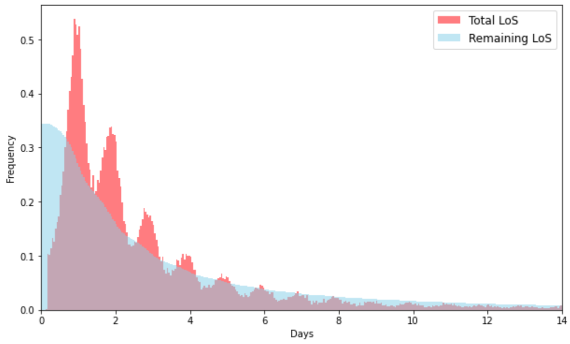

Figure 3. LoS and Remaining LoS Distributions

4.2 Dataset Description and Preprocessing

The eICU Collaborative Research Database (eICU-CRD) (Pollard et al., 2018) is a freely available, multi-center ICU database. It comprises over 200,000 patient ICU encounters for 139,367 unique patients admitted between 2014 and 2015. Patients were admitted to one of the 335 units at 208 hospitals located throughout the United States. It is a collection of a number of tables and the tables are all linked by a set of identifiers, such as *patientunitstayid* which uniquely identifies a single ICU stay, and *hospitalid* which uniquely identifies a hospital (Pollard et al., 2018).

As patient outcomes are influenced by a broad range of factors ranging from hospital characteristics (e.g., size and geographical location) to types of medical care each hospital provides, and the way in which a hospital appropriates its investments in technologies such as EHRs (Agarwal et al., 2010; Y.-K. Lin, M. Lin, and H. Chen, 2019), we consider two major kinds of modalities as the inputs for LOS prediction—static tabular data and time-series data. We then introduce all kinds of features we consider in this study and go through the preprocessing steps for each of them.

Static tabular data: To obtain static features from tabular data, various tables from the original database are utilized, such as *patient* and *hospital*, as well as several tables related to APACHE predictions. These static features encompass a range of patient and hospital attributes, including gender, age, admission and discharge dates, and regional information. In total, 17 static features are selected and subject to feature engineering, such as scaling numerical variables between -1 and 1 and converting categorical variables to one-hot encoding. Adult patients who have spent at least 5 hours in the ICU and have at least one recorded observation are included in the analysis. Furthermore, active diseases documented in the *pasthistory*, *admissiondx*, and *diagnoses* tables are extracted and represented through binary encoding. To maintain the hierarchical structure of diagnosis coding, separate features are assigned to each hierarchical level using binary encoding, as suggested by Rocheteau, Liò, and Hyland (2021). Only diagnoses that are recorded before the fifth hour in the ICU are included to prevent future data leakage. Each patient admission is represented by a single diagnosis encoding.

Time-series data: For each admission (ICU stay), 87 time-series features from the following tables: *lab*, *nursecharting*, *respiratorycharting*, *vitalperiodic*, and *vitalaperiodic*, are extracted for every hour of the ICU stay, from 24 hours before the ICU visit and up to the discharge time. A LOS target (subtracting time elapsed from the total LOS) is assigned to each timepoint in a sliding window fashion (i.e., 24 hours), starting at 5 hours after admission and ending at discharge time. In this case, the target label

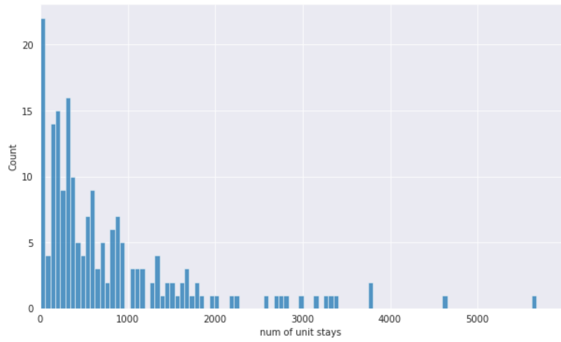

Figure 4. ICU Stays Distributions in eICU-CRD

would indicate the remaining LOS in terms of days. Moreover, time-series features are then re-sampled according to one-hour intervals and then forward-filled over the gaps to cope with missing data, because lab variables are sparsely sampled features. Any data recorded before the ICU admission will be removed after forward-filling is complete. Then, corresponding decay indicators of time-series features are added to specify how recently the observation was recorded, similar to the masking used by Che et al. (2018). Unlike static features and diagnosis coding, a sequence of time-series features is associated with each ICU admission, whose length depends on the total LOS.

To simulate a federated setting, we partition the whole database into different hospitals according to the hospital IDs to simulate a real-world FL environment in our experiments. Since the smallest unit of information is a single ICU stay, we plot the distributions of the number of ICU stays among all hospitals, as shown in Figure 4. We restrict our dataset to correspond only to hospitals whose number of ICU stays is larger than 1,000, and form a subset dataset consisting of data from 47 hospitals. The summaries of our dataset are shown in Table 1. Patient data from each hospital are randomly split into 70% for training, 15% for validation, and 15% for testing.

Description	**Value**
Number of patients	76,632
Number of unit stays	94,126
Number of hospitals	47
LoS (mean)	3.17
LoS (median)	1.90
Remaining LoS (mean)	3.99
Remaining LoS (median)	1.96
Number of input features	104
Time series features	87
Static features	17

Table 1. Subset Dataset Summaries

5 Evaluation

We evaluate our method based on the real-world eICU dataset and compared our multimodal federated method with the baselines briefly described below:

- **Centralized** learning, where a large amount of data are collected in a centralized cloud server to train a satisfactory model.
- **Local** training, where each client conducts local training using its own data without communicating with the central server using FL.

Six metrics are used to evaluate the performance: Mean Squared Logarithmic Error (MSLE), Mean Squared Error (MSE), Mean Absolute Deviation (MAD), Mean Absolute Percentage Error (MAPE), Coefficient of Determination (R^2), and Cohen Kappa Score. For MSLE, MSE, MAD, and MAPE, lower is better, and for the other two metrics, higher is better. Note that Cohen's Kappa Score (Cohen, 1960) is intended for ordered classification tasks, but it can effectively mitigate the skew if the LOS bins are chosen well. We used the same bins as (Harutyunyan et al., 2019). It is important to use multiple evaluation metrics because different metrics can give us different insights into our model's errors and fitness.

In the first part of the evaluation, we demonstrate the necessity of involving different modalities by comparing the performance of the optimized model trained with all modalities and only time-series data under the centralized setting. Based on the results presented in Table 2, it is evident that incorporating multiple modalities can enhance model performance to some degree, although this may vary depending on the tasks and the selection of modalities. For the prediction of LOS, time-series data are the most crucial modality as they align more closely with the realities of LOS prediction in clinical practice. Solely relying on tabular data will not lead to accurate predictions. However, this does not imply that tabular data, including attributes like gender, age, and ethnicity, are no longer relevant. Such modalities capture certain health discrimination patterns, which are instrumental in identifying health inequities arising from biased AI systems.

The objective of the second part is to demonstrate the effectiveness of the FL-based privacy-preserving approach. We train the TPC model using three methods: federated, centralized, and local training, on our subset dataset. In this part, we utilize all modalities and concentrate solely on comparing the differences between the three methods based on the previous results indicating that incorporating all modalities yields the best performance (see last three rows of Table 2). We evaluate the performance by employing the best models obtained during the training phase for each method and testing them on the testing data. For the federated learning approach, we need to specify two significant hyperparameters - the number of communication rounds and the number of local epochs, in addition to the common hyperparameters in conventional ML settings, to acquire the optimized trained model. Since we use FedAvg, we are required to choose a proportion of hospitals for each round of the FL training. Given that the number of hospitals in our subset dataset is not substantial, we do not perform sampling for each training round, and all hospitals conduct local training in every communication round. The local epochs and total communication rounds, on the other hand, usually need to be fine-tuned in conjunction with varying combinations, such as 1 & 100, 5 & 20, etc. It is worth noting that for large numbers of local epochs, FedAvg may not perform optimally (McMahan et al., 2017). Given the limited number of patient data, we set the local epochs to 1 and the communication rounds to 100, which proved to be sufficient for our dataset. Subsequently, we perform a search for the learning rate and batch size with a fixed number of local epochs. When it comes to centralized learning and local training, we utilize the same hyperparameter tuning process as we typically do when training ML models.

After obtaining the trained model under each setting, we evaluate the models using the testing data from all hospitals. The results for the federated and local settings are obtained by averaging the testing results from each hospital. The evaluation results are shown in Table 2. The results suggest that local training yielded the poorest performance out of the three methods. This can be attributed to the insufficient and restricted data available in each hospital, making it highly challenging to achieve optimal model training through local efforts alone without cooperation. Regarding federated and centralized learning, the findings indicate that centralized learning tends to outperform federated learning. This is because, in centralized learning, all data are aggregated and trained on a substantial amount of data. However, federated learning exhibits strong performance, with a 20% difference in MSLE compared to centralized learning — a marked improvement over local training, which is twice worse than centralized learning.

Moreover, we evaluate the performance of the models trained under the three settings in a more realistic scenario by reporting the test results of each hospital when using its own data as the test set. Interestingly, we find that the centralized trained model does not consistently outperform the federated trained model. Table 3 shows the results of the centralized trained model and its corresponding federated

Setting	MSLE	MSE	MAD	MAPE	R^2	Kappa
Centralized (only tabular)	1.99 ± 0.00	35.79 ± 0.12	3.05 ± 0.00	172.95 ± 1.18	-0.07 ± 0.01	0.10 ± 0.00
Centralized (only time-series)	0.59 ± 0.03	24.31 ± 4.45	2.07 ± 0.17	64.54 ± 5.75	0.28 ± 0.13	$\mathbf{0.66 \pm 0.00}$
Centralized (all modalities)	$\mathbf{0.59 \pm 0.01}^{**}$	$\mathbf{20.81 \pm 0.27}^{*}$	$\mathbf{1.93 \pm 0.02}$	$\mathbf{62.43 \pm 1.71}^{*}$	0.38 ± 0.01	$\mathbf{0.66 \pm 0.01}^{*}$
Local	1.90 ± 0.02	32.16 ± 3.86	2.96 ± 0.04	176.39 ± 46.17	-0.04 ± 0.09	0.16 ± 0.13
Federated	0.71 ± 0.01	22.37 ± 0.20	$\mathbf{1.93 \pm 0.01}$	72.53 ± 2.13	$\mathbf{0.41 \pm 0.00}$	0.61 ± 0.00

*Table 2. Average prediction performance of Centralized, Local, and Federated training on testing data. For each metric, the margin of error is 95% confidence intervals calculated over 3 runs. The best results are bold. Note: * p<0.05, ** p<0.01; MSLE: Mean Squared Logarithmic Error; MSE: Mean Squared Error; MAD: Mean Absolute Deviation; MAPE: Mean Absolute Percentage Error; R^2: Coefficient of Determination; Kappa: Cohen Kappa Score.*

and locally trained models from two hospitals, namely the worst-performing and best-performing hospitals. It is evident that hospitals can gain advantages from collaboration in the federated setting in contrast to local training. We observe that centralized learning outperforms federated learning in the best-performing hospital. However, centralized learning exhibits poor performance in the worst-performing hospitals, demonstrating a 40% higher MSLE than federated learning. The aforementioned observations suggest that centralized learning may exacerbate existing health disparities easily due to issues like sampling biases and the absence of representative datasets, resulting from privacy and data governance concerns. In contrast, the federated setting holds the potential to incorporate more diverse data than centralized learning in real-world scenarios, even when privacy and data governance issues persist.

Hospital	Setting	MSLE	MSE	MAD	MAPE	R^2	Kappa
Worst	Centralized	1.06 ± 0.11	216.81 ± 8.93	4.40 ± 0.11	97.56 ± 10.62	$\mathbf{0.19 \pm 0.03}$	0.60 ± 0.05
	Local	2.16 ± 0.27	82.97 ± 7.36	5.94 ± 0.25	119.80 ± 7.31	-0.11 ± 0.10	0.34 ± 0.08
	Federated	$\mathbf{0.76 \pm 0.02}^{***}$	$\mathbf{64.24 \pm 1.11}^{***}$	$\mathbf{2.56 \pm 0.04}^{*}$	$\mathbf{69.76 \pm 1.74}^{*}$	0.17 ± 0.01	$\mathbf{0.61 \pm 0.01}$
Best	Centralized	$\mathbf{0.62 \pm 0.10}^{***}$	$\mathbf{1.54 \pm 0.34}^{***}$	$\mathbf{0.70 \pm 0.06}^{*}$	$\mathbf{54.72 \pm 4.25}$	$\mathbf{0.42 \pm 0.13}$	0.53 ± 0.10
	Local	1.81 ± 0.06	264.44 ± 4.74	4.95 ± 0.03	135.16 ± 12.24	0.01 ± 0.02	0.40 ± 0.01
	Federated	0.68 ± 0.04	30.79 ± 3.87	2.20 ± 0.09	66.33 ± 2.35	0.37 ± 0.08	$\mathbf{0.63 \pm 0.02}$

*Table 3. Comparison of prediction performance for worst-performing and best-performing hospitals. For each metric, the margin of error is 95% confidence intervals calculated over 3 runs. The best results of each hospital are bold. Note: * p<0.05, ** p<0.01, *** p<0.001.*

6 Discussion

6.1 Practical and theoretical implications

This work looks into a new opportunity arising from AI-based health systems and shows that leveraging multimodal federated learning can lead to effective and accurate LOS prediction where a patient's multimodal health information does not need to be exchanged or shared between hospitals.

Our study offers important theoretical and practical implications for different stakeholders. With large-scale multimodal federated healthcare systems, patients can receive clinical support irrespective of their treatment location, place of residence, or any other social determinants of health. This means that even patients residing in deprived areas or coming from disadvantaged communities can avail themselves of the same high-quality ML-aided clinical support that is accessible to other communities. For clinicians, their judgments might also be biased as they are usually exposed to subpopulations based on their location and demographic environment. By training multimodal predictive systems in a federated fashion, clinicians from different healthcare organizations can leverage the collective knowledge and expertise of different institutions, resulting in less biased decision-making. For hospitals and healthcare providers, by establishing multimodal federated-based systems, hospitals with different population groups can collaborate without worrying about becoming a data donor, and they can maintain full control and possession of their patient data, with complete traceability of data access and limited risk of misuse by third parties. This can help address privacy concerns that may arise from sharing sensitive patient data. Moreover, by using FL, healthcare organizations can potentially reduce the costs associated with data storage and processing, as they don't need to centralize all the data from different institutions. Hence, establishing multimodal federated-based systems can lead to better patient outcomes, reduced healthcare costs, and improved resource allocation. Although our overall system is designed for real-time LOS prediction, it can easily be generalized to other clinical tasks since the general structure of the systems should be similar.

Besides, the accuracy of LOS predictions is crucial for effective patient care and resource management in healthcare organizations. However, the existing literature on LOS research has several limitations that hinder the usefulness of these predictions. There is a significant gap in the development of effective and generalized approaches for real-world LOS prediction systems that can address privacy and data governance challenges while utilizing data from multiple sources. To address these limitations, we have developed a data-driven LOS prediction system that incorporates time-series data and leverages multiple data modalities to enhance prediction accuracy. We have evaluated its effectiveness and yielded promising results that indicate the potential of our approach to enhance patient outcomes and healthcare delivery. Our approach represents a significant improvement over existing methods and can be classified as an innovative contribution to the knowledge base according to Gregor and Hevner (2013). FL offers several benefits, such as enabling the training of ML models on large datasets while keeping data distributed, without requiring centralized data storage or processing power. This can enhance privacy and security by minimizing the risk of sensitive data leakage during data sharing, thereby addressing growing concerns around data privacy and security in the IS field. Additionally, FL-based systems can reduce the computational load on centralized servers, allowing for the training of complex models on large datasets without requiring excessive resources, which can tackle the challenges of scaling up ML models to handle massive datasets. Moreover, FL can facilitate decision-making in scenarios where quick decisions are necessary and centralized decision-making is not practical, which can help overcome the obstacles of efficient decision-making in the IS field. In addition, FL can also enable personalized healthcare solutions, which can improve the accuracy of clinical prediction models and enable healthcare providers to make more informed decisions about patient care. This is related to personalized healthcare in IS research. Therefore, relevance and rigor of our proposed design ensure a grounded contribution to the knowledge base that fits inside the privacy IS, AI-related research, and healthcare domain.

6.2 Limitations and Future Research

Despite the above implications for different stakeholders, we are aware of some potential limitations of our work. Also, our findings have implications for future research. Therefore, we now outline these limitations and opportunities for future research. First, we only apply a standard FL algorithm in our proposed method, which can only learn a global model which will be used by all organizations. However, in many cases, integrating two dissimilar datasets may result in worse performance of the model trained on each local dataset. This is especially true in the medical domain, where patient data from different hospitals may exhibit variations from the general population's distribution. As a result, there is a need for additional design for FL algorithms. One avenue for future research is to incorporate personalization into the learning process rather than attempting to arrive at a single consensus model. This approach involves training multiple models for different hospitals, each tailored to the unique characteristics of that hospital. By doing so, we can better address the challenges that arise from non-IID data distributions. Furthermore, a promising direction for future research is the integration of fairness-aware algorithms. This is because there are still concerns regarding the potential for bias in the FL process, which can lead to disparities in model performance and even discrimination. Although improving prediction accuracy across hospitals is essential, it does not necessarily guarantee fairness across different demographic groups. The reason for this is that healthcare data is often highly sensitive and can exhibit variations across different demographic groups. Without accounting for these disparities, models trained on such data may not be accurate or equitable. In addition, FL-based approaches can offer a certain degree of privacy because the data from each institution will never be directly accessed by others, and only model parameters will be shared and aggregated on the server. However, models themselves can memorize some information under certain conditions (Sablayrolles et al., 2019), which would still cause information leakage sometimes. Therefore, further analysis on privacy leakage will be beneficial, and privacy mechanisms such as differential privacy (Abadi et al., 2016), or cryptographic methods can be added to the FL setting to further enhance privacy. Moreover, in this study, we have only used a limited number of modalities. However, there is potential for future research to explore the use of additional modalities, such as unstructured clinical notes, written reports, and even image data, to further enhance the prediction capabilities. Finally, further research can engage practice partners to implement a field test to validate the effectiveness of our multimodal federated LOS prediction system. We also plan to carry out a user test to obtain valuable insights into the perspectives of practitioners and incorporate multiple design cycles to further enhance our approach.

7 Concluding Remarks

In this study, a multimodal federated-learning-based framework is presented to predict real-time LOS in the ICUs. In the context of a real-world medical dataset, we explore the potential of multimodal federated learning in addressing the privacy and data governance issues related to healthcare and enabling the use of multimodal data in medical applications. We implemented and evaluated a multimodal federated-learning-based approach for ICU remaining LOS prediction, which is capable of predicting patient remaining LOS every hour throughout a patient's entire stay in the ICU by leveraging multimodal data and privacy-preserving federated learning. Our design has a series of practical implications for different stakeholders and shows the potential of applying large-scale multimodal federated healthcare systems to mitigate the disparities caused by privacy concerns and limited-available data in centralized ML systems. As for the theoretical implications, we addressed a relevant real-world problem with a set of design decisions and it contributes to the emerging literature on privacy IS, especially privacy research in health informatics, and sheds light on the importance and necessity of using multimodal inputs in the healthcare domain. In the future, we plan to further improve our work by involving personalized and fairness-aware FL algorithms, additional privacy mechanisms, as well as more modalities.

References

Abadi, M., A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang (2016). "Deep learning with differential privacy." In: *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pp. 308–318.

Abouelmehdi, K., A. Beni-Hessane, and H. Khaloufi (2018). "Big healthcare data: preserving security and privacy." *Journal of big data* 5 (1), 1–18.

Agarwal, R., G. Gao, C. DesRoches, and A. K. Jha (2010). "Research commentary—The digital transformation of healthcare: Current status and the road ahead." *Information systems research* 21 (4), 796–809.

Alsinglawi, B., O. Alshari, M. Alorjani, O. Mubin, F. Alnajjar, M. Novoa, and O. Darwish (2022). "An explainable machine learning framework for lung cancer hospital length of stay prediction." *Scientific reports* 12 (1), 1–10.

Anderson, C. L. and R. Agarwal (2011). "The digitization of healthcare: boundary risks, emotion, and consumer willingness to disclose personal health information." *Information Systems Research* 22 (3), 469–490.

Awad, A., M. Bader-El-Den, J. McNicholas, and J. Briggs (2017). "Early hospital mortality prediction of intensive care unit patients using an ensemble learning approach." *International journal of medical informatics* 108, 185–195.

Bacchi, S., Y. Tan, L. Oakden-Rayner, J. Jannes, T. Kleinig, and S. Koblar (2022). "Machine learning in the prediction of medical inpatient length of stay." *Internal medicine journal* 52 (2), 176–185.

Baek, H., M. Cho, S. Kim, H. Hwang, M. Song, and S. Yoo (2018). "Analysis of length of hospital stay using electronic health records: A statistical and data mining approach." *PloS one* 13 (4), e0195901.

Bartel, A. P., C. W. Chan, and S.-H. Kim (2020). "Should hospitals keep their patients longer? The role of inpatient care in reducing postdischarge mortality." *Management Science* 66 (6), 2326–2346.

Beckerman, J., J. Pritts, E. Goplerud, J. Leifer, P. Borzi, S. Rosenbaum, and D. Anderson (2008). "A delicate balance: Behavioral health, patient privacy, and the need to know." *California Healthcare Foundation: Issue Brief*, 1–12.

Brisimi, T. S., R. Chen, T. Mela, A. Olshevsky, I. C. Paschalidis, and W. Shi (2018). "Federated learning of predictive models from federated electronic health records." *International journal of medical informatics* 112, 59–67.

Che, Z., S. Purushotham, K. Cho, D. Sontag, and Y. Liu (2018). "Recurrent neural networks for multivariate time series with missing values." *Scientific reports* 8 (1), 1–12.

Clifton, C., M. Kantarcioglu, J. Vaidya, X. Lin, and M. Y. Zhu (2002). "Tools for privacy preserving distributed data mining." *ACM Sigkdd Explorations Newsletter* 4 (2), 28–34.

Daghistani, T. A., R. Elshawi, S. Sakr, A. M. Ahmed, A. Al-Thwayee, and M. H. Al-Mallah (2019). "Predictors of in-hospital length of stay among cardiac patients: a machine learning approach." *International journal of cardiology* 288, 140–147.

Davoodi, R. and M. H. Moradi (2018). "Mortality prediction in intensive care units (ICUs) using a deep rule-based fuzzy classifier." *Journal of biomedical informatics* 79, 48–59.

De Fauw, J., J. R. Ledsam, B. Romera-Paredes, S. Nikolov, N. Tomasev, S. Blackwell, H. Askham, X. Glorot, B. O'Donoghue, D. Visentin, et al. (2018). "Clinically applicable deep learning for diagnosis and referral in retinal disease." *Nature medicine* 24 (9), 1342–1350.

Department of Health and Human Services (2013). *Modifications to the HIPAA Privacy, Security, Enforcement, and Breach Notification Rules Under the Health Information Technology for Economic and Clinical Health Act and the Genetic Information Nondiscrimination Act; Other Modifications to the HIPAA Rules*. Accessed: November 3, 2022. URL: `https://www.govinfo.gov/content/pkg/FR-2013-01-25/pdf/2013-01073.pdf`.

Duan, Y., J. Canny, and J. Zhan (2010). "{P4P}: Practical {Large-Scale}{Privacy-Preserving} Distributed Computation Robust against Malicious Users." In: *19th USENIX Security Symposium (USENIX Security 10)*.

Dwork, C., F. McSherry, K. Nissim, and A. Smith (2006). "Calibrating noise to sensitivity in private data analysis." In: *Theory of cryptography conference*. Springer, pp. 265–284.

Ghassemi, M., T. Naumann, P. Schulam, A. L. Beam, I. Y. Chen, and R. Ranganath (2020). "A review of challenges and opportunities in machine learning for health." *AMIA Summits on Translational Science Proceedings* 2020, 191.

Gregor, S. and A. R. Hevner (2013). "Positioning and presenting design science research for maximum impact." *MIS quarterly*, 337–355.

Gunčar, G., M. Kukar, M. Notar, M. Brvar, P. Černelč, M. Notar, and M. Notar (2018). "An application of machine learning to haematological diagnosis." *Scientific reports* 8 (1), 1–12.

Hachesu, P. R., M. Ahmadi, S. Alizadeh, and F. Sadoughi (2013). "Use of data mining techniques to determine and predict length of stay of cardiac patients." *Healthcare informatics research* 19 (2), 121–129.

Harutyunyan, H., H. Khachatrian, D. C. Kale, G. Ver Steeg, and A. Galstyan (2019). "Multitask learning and benchmarking with clinical time series data." *Scientific data* 6 (1), 1–18.

Hevner, A. R., S. T. March, J. Park, and S. Ram (2004). "Design science in information systems research." *MIS quarterly*, 75–105.

Huang, L., A. L. Shea, H. Qian, A. Masurkar, H. Deng, and D. Liu (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.

Huang, S.-C., A. Pareek, S. Seyyedi, I. Banerjee, and M. P. Lungren (2020). "Fusion of medical imaging and electronic health records using deep learning: a systematic review and implementation guidelines." *NPJ digital medicine* 3 (1), 1–9.

Kalchbrenner, N., L. Espeholt, K. Simonyan, A. v. d. Oord, A. Graves, and K. Kavukcuoglu (2016). "Neural machine translation in linear time." *arXiv preprint arXiv:1610.10099*.

Laur, S., H. Lipmaa, and T. Mielikäinen (2006). "Cryptographically private support vector machines." In: *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 618–624.

Lee, G. and S.-Y. Shin (2020). "Federated learning on clinical benchmark data: performance assessment." *Journal of medical Internet research* 22 (10), e20891.

Lee, J., J. Sun, F. Wang, S. Wang, C.-H. Jun, X. Jiang, et al. (2018). "Privacy-preserving patient similarity learning in a federated environment: development and analysis." *JMIR medical informatics* 6 (2), e7744.

Leslie, D., A. Mazumder, A. Peppin, M. K. Wolters, and A. Hagerty (2021). "Does "AI" stand for augmenting inequality in the era of covid-19 healthcare?" *bmj* 372.

Lin, Y.-K., M. Lin, and H. Chen (2019). "Do electronic health records affect quality of care? Evidence from the HITECH Act." *Information Systems Research* 30 (1), 306–318.

Lin, M., Q. Chen, and S. Yan (2013). "Network in network." *arXiv preprint arXiv:1312.4400*.

Liu, B., P. A. Pavlou, and X. Cheng (2022). "Achieving a Balance Between Privacy Protection and Data Collection: A Field Experimental Examination of a Theory-Driven Information Technology Solution." *Information Systems Research* 33 (1), 203–223.

McMahan, B., E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas (2017). "Communication-efficient learning of deep networks from decentralized data." In: *Artificial intelligence and statistics*. PMLR, pp. 1273–1282.

Morton, A., E. Marzban, G. Giannoulis, A. Patel, R. Aparasu, and I. A. Kakadiaris (2014). "A comparison of supervised machine learning techniques for predicting short-term in-hospital length of stay among diabetic patients." In: *2014 13th International Conference on Machine Learning and Applications*. IEEE, pp. 428–431.

Myszczynska, M. A., P. N. Ojamies, A. Lacoste, D. Neil, A. Saffari, R. Mead, G. M. Hautbergue, J. D. Holbrook, and L. Ferraiuolo (2020). "Applications of machine learning to diagnosis and treatment of neurodegenerative diseases." *Nature Reviews Neurology* 16 (8), 440–456.

Obermeyer, Z., B. Powers, C. Vogeli, and S. Mullainathan (2019). "Dissecting racial bias in an algorithm used to manage the health of populations." *Science* 366 (6464), 447–453.

Oord, A. v. d., S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu (2016). "Wavenet: A generative model for raw audio." *arXiv preprint arXiv:1609.03499*.

Panch, T., H. Mattie, and R. Atun (2019). "Artificial intelligence and algorithmic bias: implications for health systems." *Journal of global health* 9 (2).

Panch, T., H. Mattie, and L. A. Celi (2019). "The "inconvenient truth" about AI in healthcare." *NPJ digital medicine* 2 (1), 1–3.

Peffers, K., T. Tuunanen, M. A. Rothenberger, and S. Chatterjee (2007). "A design science research methodology for information systems research." *Journal of management information systems* 24 (3), 45–77.

Pfohl, S. R., A. M. Dai, and K. Heller (2019). "Federated and differentially private learning for electronic health records." *arXiv preprint arXiv:1911.05861*.

Pollard, T. J., A. E. Johnson, J. D. Raffa, L. A. Celi, R. G. Mark, and O. Badawi (2018). "The eICU Collaborative Research Database, a freely available multi-center database for critical care research." *Scientific data* 5 (1), 1–13.

Rajkomar, A., E. Oren, K. Chen, A. M. Dai, N. Hajaj, M. Hardt, P. J. Liu, X. Liu, J. Marcus, M. Sun, et al. (2018). "Scalable and accurate deep learning with electronic health records." *NPJ digital medicine* 1 (1), 1–10.

Rieke, N., J. Hancox, W. Li, F. Milletari, H. R. Roth, S. Albarqouni, S. Bakas, M. N. Galtier, B. A. Landman, K. Maier-Hein, et al. (2020). "The future of digital health with federated learning." *NPJ digital medicine* 3 (1), 1–7.

Rocheteau, E., P. Liò, and S. Hyland (2021). "Temporal pointwise convolutional networks for length of stay prediction in the intensive care unit." In: *Proceedings of the Conference on Health, Inference, and Learning*, pp. 58–68.

Rotter, T., L. Kinsman, E. L. James, A. Machotta, H. Gothe, J. Willis, P. Snow, and J. Kugler (2010). "Clinical pathways: effects on professional practice, patient outcomes, length of stay and hospital costs." *Cochrane database of systematic reviews* (3).

Sablayrolles, A., M. Douze, C. Schmid, Y. Ollivier, and H. Jégou (2019). "White-box vs black-box: Bayes optimal strategies for membership inference." In: *International Conference on Machine Learning*. PMLR, pp. 5558–5567.

Sedayao, J., R. Bhardwaj, and N. Gorade (2014). "Making big data, privacy, and anonymization work together in the enterprise: experiences and issues." In: *2014 IEEE International Congress on Big Data*. IEEE, pp. 601–607.

Soenksen, L. R., Y. Ma, C. Zeng, L. Boussioux, K. Villalobos Carballo, L. Na, H. M. Wiberg, M. L. Li, I. Fuentes, and D. Bertsimas (2022). "Integrated multimodal artificial intelligence framework for healthcare applications." *NPJ digital medicine* 5 (1), 1–10.

Song, H., D. Rajan, J. Thiagarajan, and A. Spanias (2018). "Attend and diagnose: Clinical time series analysis using attention models." In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 32. 1.

Sotoodeh, M. and J. C. Ho (2019). "Improving length of stay prediction using a hidden Markov model." *AMIA Summits on Translational Science Proceedings* 2019, 425.

Sweeney, L. (2000). "Simple demographics often identify people uniquely." *Health (San Francisco)* 671 (2000), 1–34.

Trumbo, C. W., K. A. McComas, and P. Kannaovakun (2007). "Cancer anxiety and the perception of risk in alarmed communities." *Risk Analysis: An International Journal* 27 (2), 337–350.

Tsai, P.-F. J., P.-C. Chen, Y.-Y. Chen, H.-Y. Song, H.-M. Lin, F.-M. Lin, and Q.-P. Huang (2016). "Length of hospital stay prediction at the admission stage for cardiology patients using artificial neural network." *Journal of healthcare engineering* 2016.

Vaid, A., S. K. Jaladanki, J. Xu, S. Teng, A. Kumar, S. Lee, S. Somani, I. Paranjpe, J. K. De Freitas, T. Wanyan, et al. (2021). "Federated learning of electronic health records to improve mortality prediction in hospitalized patients with COVID-19: machine learning approach." *JMIR medical informatics* 9 (1), e24207.

Voigt, P. and A. Von dem Bussche (2017). "The eu general data protection regulation (gdpr)." *A Practical Guide, 1st Ed., Cham: Springer International Publishing* 10 (3152676), 10–5555.

Wang, F., L. P. Casalino, and D. Khullar (2019). "Deep learning in medicine—promise, progress, and challenges." *JAMA internal medicine* 179 (3), 293–294.

Xu, J., B. S. Glicksberg, C. Su, P. Walker, J. Bian, and F. Wang (2021). "Federated learning for healthcare informatics." *Journal of Healthcare Informatics Research* 5 (1), 1–19.