

MTFed: Multitask Clinical Representation Learning via Multi-Institutional Federation with EMR Privacy Preserving

Junyi Gao

junyi5, Computer Science, UIUC
https://youtu.be/eufoWbN_3SY

ABSTRACT

Deep learning-based health status representation learning is a fundamental research problem for clinical prediction (e.g., the prognosis of COVID-19) and has received a huge amount of attention from the community. The training of deep models requires a large amount of high-quality data. However, Electronic Medical Records are usually difficult to obtain timely due to the lack of accurate diagnostic methods and privacy concerns, especially at the early stage of an emerging pandemic. In practise, hospitals usually do not share any common samples and, at most of the time, only share part of medical features. The expected prediction targets also differ among institutions. This presents a challenging normative question: *How to effectively connect these institutions to improve the performance of different targets and perform clustering study without sharing the patient data?* In this paper, we propose a *multitask* clinical representation learning framework, called MTFed. It embeds feature sequences separately based on a multi-channel structure. Collaborators can share the low-level feature extraction layers to obtain the robust embedding jointly. We also develop a feature calibration module, which can avoid the distraction of unrecorded or useless features from each collaborator. In MTFed, federated clustering is also adopted to assist physicians to perform cohort study while preserving the privacy of the patients. Our extensive experimental results demonstrate the effectiveness of MTFed. To facilitate the personalized clinical service and verify the reasonability of the model, we also develop a real-world AI-Doctor interaction system to dynamically visualize the patient's health trajectory.

CCS CONCEPTS

• Information systems → Data mining; • Applied computing → Health informatics.

KEYWORDS

Federated Learning, COVID-19, Privacy Preserving, Medical Informatics

ACM Reference Format:

Junyi Gao. 2021. MTFed: Multitask Clinical Representation Learning via Multi-Institutional Federation with EMR Privacy Preserving. In *Ljubljana*

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

Ljubljana '21, April, 2021, Ljubljana, Slovenia

© 2021 Association for Computing Machinery.

ACM ISBN 978-1-4503-XXXX-X/18/06...\$15.00

<https://doi.org/10.1145/1122445.1122456>

'21: *The Web Conference, April, 2021, Ljubljana, Slovenia*. ACM, New York, NY, USA, 8 pages. <https://doi.org/10.1145/1122445.1122456>

1 INTRODUCTION

The sudden outbreak of the epidemic COVID-19 has caused continuous damage to the world. For patients with COVID-19 who face severe life threats and receive ICU treatments, their health conditions are continually changing over time. Physicians need to evaluate patients' health, then select personalized follow-up treatments to prevent adverse outcomes and assign medical resources effectively, as well as reduce medical costs.

Recently the AI-based clinical healthcare prediction on COVID-19 Electronic Medical Record (EMR) has raised significant research interest from the community. Previous works [36] indicate that applying machine learning to healthcare prediction can improve the prognostic effect and optimize the utilization of medical resources during the pandemic. Usually, those EMR analysis models first embed multiple dynamic features (e.g., lab test values) into a low-dimensional feature space for each clinical visit, then learn the dense embedding of the patient's health status through the entire visits by sequential models, and finally perform specific clinical analysis tasks based on the learned representation.

However, to be practically deployed in medical scenarios, AI algorithms need to reach clinical-grade accuracy. As such, the models usually need a large and balanced labeled dataset for training. However, in practice, AI practitioners are facing the below challenges in related clinical applications.

- **Data volume insufficiency:** Training the model on relatively small local labeled datasets will potentially result in the under-performance of deep learning models due to over-fitting. The robustness of the model would also be damaged due to the data insufficiency [17].
- **Data diversity insufficiency:** Training examples need to sufficiently represent the clinical environment in which they will be used. Besides, physicians urgently need to perform cohort analysis for patients in diverse conditions reasonably. However, the data collected by a single medical institution can be biased by, for example, patient demographics, the instruments used or clinical specializations. When considering rare disease conditions of some rare patients, there may be only a few cases of such a particular condition. The reasonability of cohort analysis would be jeopardized, leading to erroneous decisions in clinical practice.

But the data (e.g., lab tests, clinical outcomes) have traditionally proved hard to accumulate, especially at the early stage of an emerging pandemic like COVID-19. The reasons can be summarized as follows: 1) The precise diagnostic mechanism was not established in the early outbreak. Before introducing the nucleic acid detection

mechanism, it is difficult to confirm whether a patient is really infected with COVID-19, so that researchers may not acquire enough labeled data. For example, there are only 41 patients diagnosed with COVID-19 due to the lack of valid testing methods at the early outbreak in Wuhan, [16]. 2) The disease is still progressing for patients. The collection of enough outcomes needs to take a long time. 3) The issue of clinical data scarcity would be further exacerbated due to the missing or inconsistent follow-up clinical visits.

As a result, integrating the EMR from different institutions to augment the training data is urgently needed. It also needs to be very diverse and incorporate balanced data from patients of different genders, ages, demographics and environmental exposure. Nonetheless, integrating a sufficiently large and balanced dataset from various institutions is a major medical challenge. Specifically, how to integrate the institutions with various recorded clinical feature sets and expected clinical prediction targets to *jointly improve the performance* of their tasks, when *assuring privacy protection* through the training procedure? The issues that need to be tackled are summarized as follows:

- **I1: Integration of institutions with various recorded clinical feature sets.** The clinical institutions usually differ in feature spaces, due to the limitation of equipment or the different medical practice guidelines. For example, *Tongji Hospital* in China has recorded 74 sequential clinical features (e.g., *Serum Chloride*, *Hemoglobin*) for COVID-19 patients. There are only 27 common features (e.g., *Urea*) shared by *Tongji Hospital* and *HM Hospitals* in Spain [15, 36]. Furthermore, they also have very different patient groups from their service regions, and the intersection set of their patients is usually very small, unlike most vertical federated learning situations. Thus, it is critical to make full use of the various medical features to represent patients' health status when integrating into clinical institutions.
- **I2: Integration of institutions with various expected clinical prediction targets.** The clinical prediction target in critical needs may differ among institutions. For example, evaluating patients' health risk and predicting the remaining in-hospital period helps to assign medical resources effectively, especially when the hospitals are overwhelmed by the sudden emerging pandemics such as COVID-19. On the other hand, early detection of sepsis is expected to improve outcomes in the *PhysioNet-Cardiology* institution, America. It is believed that fully exploring the useful information stored in various targets will enhance the overall health representation learning performance.
- **I3: Serious privacy concerns about individual clinical data.** Multi-agency collaboration based on centralized shared data faces privacy and ownership challenges. To protect patients' privacy, sharing medical data, such as the raw dynamic sequential lab tests and the individual demographic information, across medical institutions is usually not allowed. The data-sharing mechanism across multiple hospitals worldwide usually cannot be established timely, especially for the emerging pandemic. As a result, it is critical to jointly build a federated model without sharing raw patient data.

In this paper, we propose a multitask health representation learning framework, MTFed, based on the multi-institutional federation. This approach enables several organizations to collaborate on the

development of models, but without the need to directly share any local sensitive raw data among each other. In each round of training, collaborators (e.g., hospitals) are selected to train a model using local data. Only model updates are sent to the central server for aggregation to preserve data privacy. Concretely, multi-channel architecture is utilized to embed each clinical feature separately in the clinical representation learning to improve the compatibility across institutions with different feature sets. Each collaborator can borrow useful feature extractors trained by other institutions even with different prediction targets, thus reducing the difficulty of its task-specific prediction layer. Feature-wise recalibration is deployed to further adaptively emphasize critical features for various tasks. Finally, patients in similar health conditions but different hospitals are clustered into federated groups across institutions to help physicians reasonably perform cohort study for specific patient groups. Our contributions are summarized as follows:

- We propose a federated clinical prediction framework, MTFed, which builds robust health representation via getting exposed to a significantly wider range of data from institutions with various prediction tasks. MTFed helps to improve the performance of health evaluation (e.g., the prognosis of COVID-19 patients) effectively for all collaborators under the situation of data volume insufficiency, especially at the early stage of an emerging pandemic.
- We design a federated clustering analysis method, which combines patients in hospitals with various feature sets to build a federation dataset with wider demographic ranges and increase patient conditions' data diversity. It helps physicians evaluate the health status in a reasonable clustering perspective for a particular clinical area or rare cases that they would not have come across locally. They can also perform a comprehensive cohort study for patients in similar conditions but different hospitals.
- The proposed decentralized training framework, MTFed, avoids locally aggregating raw clinical data across multiple institutions, and thus preserves the privacy of the individual clinical data. This helps to accelerate the model building procedure and relate medical studies of an emerging pandemic.
- We connect isolated medical institutions (i.e., *Tongji Hospital* in China, *HM Hospitals* in Spain, *PhysioNet Institution* in America) and train models for various prediction tasks (i.e., Length of stay prediction of COVID-19, Sepsis prediction of Cardiology) in a way that preserves data privacy. Experiment results show that the proposed federated framework jointly improves the prediction performance for all collaborators.

2 RELATED WORK

2.1 Clinical Background: COVID-19

Outbreaks of the COVID-19 epidemic have been arising worldwide health concerns and were declared a pandemic by the World Health Organization (WHO). Although the huge impact of COVID-19 is uncertain, it has significantly overwhelmed health care infrastructure. All emerging viral pandemics can place extraordinary and sustained demands on public health systems and essential community service providers [26]. Limited healthcare resource availability will increase the chance of being infected while waiting for treatment and mortality rates [18]. This eventually leads to an increase

in the severity of the pandemic. With the continuous development of the new coronavirus epidemic, more and more researchers are committed to joining the ranks of fighting the epidemic through AI-related technologies. Massive COVID-related research works focus on the severity of disease rather than the clinical outcome of mortality [7, 10, 35]. These studies answer critical clinical questions on COVID-19 evolution and outcomes, as well as potential risk factors leading to hospital and ICU admission. However, they cannot make individualized risk predictions for patients. Recently, Li et al. [36] uses machine learning-based methods such as decision trees to make risk predictions for COVID-19 patients. To optimize patient care and appropriately deploy healthcare resources during this pandemic, effective and reliable early risk prediction is still a crucial and urgent problem.

2.2 Deep-Learning-Based EMR Analysis

With the development of healthcare together with the update of storage, there are much valuable digital information stored in electronic medical records (EMR), which opens a door for researchers to make secondary use of these records for various clinical applications [8, 19, 37, 39]. Many deep learning-based models have been developed to mine the massive EMR data due to the remarkable representation learning ability of neural networks. Existing methods have shown superior performance in many tasks, such as mortality prediction [9, 14, 33], patients subtyping [1], and diagnosis prediction [1, 4, 20, 24, 27, 29]. Though the medical tasks vary from each other, extracting advanced clinical features and learning the compressed representation of the sparse EMR data are fundamental procedures of clinical healthcare prediction. Such representations can characterize patients' information in low-dimensional space, thus making the mortality risk and disease diagnosis prediction easier.

However, training deep-learning-based models usually needs a large amount of data with high diversity, representing the practical application environment. The quantity of labeled data is much less for some rare diseases, which cannot support a model to be trained thoroughly.

Recently some researchers try to exploit additional information to deal with the scarcity of clinical data. For example, [13] trains a model with multiple related tasks (e.g., mortality prediction, phenotyping, length of stay). But this needs extra labeling efforts from human medical experts. [5] leverages the inherent multilevel structure (e.g., the relationship between diagnosis codes and treatment codes) of EHR data to improve learning efficiency. [3] and [22] introduce external well-organized ontology information (e.g., International Classification of Diseases Codes) to represent the medical concept as a combination of its ancestors in the ontology via an attention mechanism. However, such relationships and ontology information are often not easy to access in clinical practice. Besides, ontological information is usually designed to handle the medical codes. Thus it is not suitable for dealing with numerical lab tests, which also are essential clinical features to capture health status. For example, there is not a kind of normal structured information of relationship information among lab test values (e.g., blood glucose, hemoglobin).

On the other hand, some researchers try to explore the existing EMR data. Choi [2] empirically confirms that RNN models possess great potential for transferring learning across different medical institutions. Gupta [11] trains a deep RNN to identify several patient phenotypes on time series from MIMIC-III dataset, and then uses the features extracted by the RNN to build classifiers for identifying previously unseen phenotypes. However, these methods can only be utilized for the same tasks with the same clinical feature sets between source and target datasets. TimeNet [12] is pre-trained on non-medical time series in an unsupervised manner and further utilized to extract features for clinical prediction. Nevertheless, the trained parameters for the non-medical data may not be suitable for the specific clinical task, leading to negative transfer and limited performance.

2.3 Federated Learning for Healthcare

Federated learning is recent emerging research that has been extensively studied in the fields of financial security. It is a novel paradigm of multi-institutional collaboration for data privacy, in which model learning uses all available data without sharing data between institutions. Collaborators train a shared global model with a server across multiple decentralized clients holding local data samples, and then only the updated results are aggregated to the server [25]. After joint optimization, the server returns the global state to clients, and continues to accept the updated data calculated by each client in the new global state. Federated learning is suitable for solving the problem of machine learning using large-scale private data due to its security and privacy.

In the medical record analysis area, for the protection and respect of patients' privacy, the hospital's specific medical-related data did not allow leakage and sharing without permission [21]. Sheller [31, 32] introduces the first use of federated learning to perform brain tumor segmentation.

Collecting the training data of COVID-19 was a major challenge at the early stage of the emerging pandemic and has caused a lack of sufficient data samples when performing deep learning approaches [21]. Zhang [38] and Liu [21] propose federated learning approaches for medical diagnostic image (i.e., Chest X-ray Images) analysis to detect COVID-19 infections. Vaid [34] employs Logistic Regression and Multilayer Perceptron to predict the mortality for COVID-19 patients based on federated learning. However, these methods can only be applied among institutions with the same prediction target (e.g., COVID-19 detection, mortality prediction) and the same feature space (e.g., image or same clinical features recorded). This severely limits application scenarios.

3 PROBLEM FORMULATION

Medical institutions expect to build a federation together to learn a robust representation learning model and perform accurate task predictions. Concretely, define N medical collaborators (e.g., hospitals) $\{C^1, \dots, C^N\}$ owning the data $\{D^1, \dots, D^N\}$ with their clinical prediction targets $\{Y^1, \dots, Y^N\}$ respectively. Hospitals may share different clinical prediction targets. For example, many hospitalized patients with COVID-19 face severe life threats and need careful health monitoring. The COVID-19 collaborators may expect to evaluate the health status and predict remaining time spent in ICU (i.e.,

length of stay) for patients, helping assess the severity of illness and assign medical resources [28]. On the contrary, cardiology collaborators may expect to perform the early prediction of sepsis to improve the outcome of other patients. Specifically, for a given collaborator C^n :

$$D^n = \{(r_i^n, y_i^n)\}_{i=1}^{M^n}, \quad (1)$$

where r_i^n is the medical records of patient i , and M^n is the data size of collaborator C^n . Hospitals usually also record various medical feature sets (e.g., medical biomarkers, vital signs),

$$r_i^n = \begin{pmatrix} r_{1,1} & \cdots & r_{1,T} \\ \vdots & \ddots & \vdots \\ r_{x^n,1} & \cdots & r_{x^n,T} \end{pmatrix}. \quad (2)$$

where x^n is the feature size of collaborator F^n , and T is the maximum timestep size. Some collaborators may share a part of feature sets,

$$X^n \cap X^j \neq \emptyset, \quad \forall D^n, D^j, n \neq j \quad (3)$$

where X^n is the feature sets recorded in collaborator C^n . A conventional aggregating training method is to put all institutions' data with the same feature sets and the same prediction targets together to train a model. However, it is tough to satisfy this condition, especially at the early stage of the emerging epidemic. Furthermore, it will also cause unavoidable privacy leakage. Federated learning decentralizes deep learning by removing the need to pool data into a single location. Instead, the model is trained in multiple iterations at different sites. For example, say two COVID-19 institutions and one cardiology institution decide to team up, jointly building models to predict the length-of-stay of COVID-19 patients and perform the sepsis early detection of other patients. Concretely, in this paper, multitasking multi-institutional federation learning is proposed to improve the performance of the various prediction tasks for all collaborators with different feature sets. It deals with the problem exceeding the scope of basic federated learning:

$$r^n \cap r^j = \emptyset, \quad X^n \neq X^j, \quad Y^n \neq Y^j, \quad \forall D^n, D^j, n \neq j \quad (4)$$

4 METHODOLOGY

4.1 Multi-Variable Sequential Medical Records Representation Learning

In order to facilitate each hospital with different characteristics as participants to make better use of the federal learning framework, we utilize the multi-channel clinical sequence embedding [23]. Specifically, each clinical feature is embedded by RNN separately:

$$f_x = \text{GRU}_x(r_{x,1}, \dots, r_{x,T}) \quad (5)$$

Furthermore, the demographic baseline data (e.g., age, gender, primary disease) $base_1, base_2, \dots, base_b$ is embedded into the same hidden space of f_x as hidden size h , where $W^{base} \in \mathbb{R}^{m \times h}$ is a learnable embedding matrix.

$$f_{x+1} = W^{base} \cdot base \quad (6)$$

Thus, all the data of the patient can be represented by a matrix $F = (f_1, \dots, f_x, f_{x+1})^T$ which is a sequence of vectors and each vector represents one feature of the patient over time.

4.2 Multitask Multi-Institutional Federated Learning with Various Feature Sets

For the basic federated learning of EMR analysis, a centralized server would maintain the global deep neural network and each participating hospital would be given a copy to train on their own dataset. To preserve the privacy of health data, each client trains on the local model using the local dataset in each round of training, and then encrypts the updated parameters of the model and uploads it to the server. After the server receives the update parameters, it performs decryption and aggregation operations and then encrypts the model's global update parameters and sends them to each client. The updated parameters would then be shared with the participating institutes, so that they could decrypt and apply the updates to their models. This approach ensures collaborators' accuracy and privacy, utilizing a large amount of data to learn and maintain an excellent model.

During data preprocessing in this work, to keep the feature value's consistency, the data-preprocessing will undergo unified standardization operations. For common features, collaborators will share the metadata used for standardization (e.g., mean value and standard deviation of the shared features). All the collaborators embed their sequential records via the corresponding GRU channels in the federated framework. Hospitals with shared features borrow useful information from each other by jointly training the feature extractors in common.

The prediction tasks usually differ among collaborators, and thus the final prediction layer is supposed to be private for some hospitals. Even so, they still expect to share the low-level layers with other collaborators to jointly obtain robust embedding. However, the critical features required by various prediction targets are not exactly the same. Not all the layers trained in the federated framework are useful for the specific task. In order to avoid the distraction of unrecorded or useless features for each collaborator, feature recalibration is designed to automatically suppress the non-existent features for each hospital, and meanwhile adaptively enhance important features for patients in diverse health conditions.

Based on shared feature extraction channels, all collaborators with various tasks can embed patients with different recorded features in the same clinical feature space. This helps physicians to perform further cohort study of patient groups based on federated clustering analysis in the following subsection. The feature recalibration mechanism and the private prediction module guarantee the individuation of each collaborator. As a result, such a federated representation learning framework can jointly improve the prediction performance for each collaborator and meanwhile provide reasonable interpretability.

Specifically for collaborator C^n , we calculate the queries, keys and values for F obtained in the multi-variable sequence representation learning layer:

$$q_i = W_i^q \cdot \tilde{f}, \quad (7)$$

$$k_i = W_i^k \cdot \tilde{f}_i, \quad (8)$$

$$v_i = W_i^v \cdot \tilde{f}_i, \quad (9)$$

where W^q , W^k and W^v are the learnable projection matrix respectively, and i is from 1 to $n + 1$. The attention weights are calculated

as:

$$\alpha_1, \dots, \alpha_x, \alpha_{x+1} = \text{Softmax}(\zeta_1, \dots, \zeta_x, \zeta_{x+1}), \quad (10)$$

where

$$\zeta_i = \begin{cases} q_i \cdot k_i & , \text{if } r_i \text{ recorded in } C^n \\ \text{mask} & , \text{if } r_i \text{ unrecorded in } C^n \end{cases}$$

where *mask* is a negative number with large absolute value. The health status representation *s* can be obtained as:

$$s = \sum_{i=1}^N \alpha_i \cdot v_i, \quad (11)$$

Collaborators build their own prediction layer based on the jointly trained embedding module. For example, the classification task (e.g., mortality prediction, sepsis prediction) can be performed as:

$$\hat{y}_{cla} = \sigma(W_{cla} \cdot s + b_{cla}), \quad (12)$$

where W_{cla} and b_{cla} are the learnable matrix and bias term, respectively. Assuming M^n is the total number of samples of collaborator C^n , and the final loss can be denoted as binary cross-entropy loss:

$$\mathcal{L}_{cla} = -\frac{1}{S} \sum_{i=1}^S [y_{cla}^i \log(\hat{y}_{cla}^i) + (1 - y_{cla}^i) \log(1 - \hat{y}_{cla}^i)]. \quad (13)$$

And for another example, regression task, such as length-of-stay prediction which aims to predict the remaining days to outcome at each record of patients:

$$\hat{y}_{reg} = W_{reg} \cdot s + b_{reg}, \quad (14)$$

Similarly, W_{reg} and b_{reg} are learnable. The final loss can be regarded as mean squared error (MSE):

$$\mathcal{L}_{reg} = -\frac{1}{S} \sum_{i=1}^S (y_{reg}^i - \hat{y}_{reg}^i)^2. \quad (15)$$

Finally, hospitals with different tasks can choose to perform fine-tuning separately in local.

4.3 Federated Clustering

To promote comprehensive cohort study for patient groups in similar conditions and further extract useful medical knowledge for auxiliary clinical treatment, we design federated clustering methods for patients with different feature sets across various institutions. Considering privacy protection, both the raw health data and direct embedded representation results of patients cannot be shared among the federation. For example, both HM Hospital (Spain) and Tongji Hospital (China) are COVID-19 collaborators with the same length-of-stay prediction tasks. The recorded medical features in these institutions are not the same¹, but the health status of patients can still be embedded into the same feature space by MTFed. Then, all patients' representation results in these collaborators are clustered based on K-means with privacy preservation. Concretely, the server randomly initializes the center of clusters. After encrypting and sending the center of clusters to the server, collaborators calculate distances to centers with patients' representation, encrypt and send back to the server. The server decrypts and recalculated clusters' center until convergence. The hyper-parameter *k* in the clustering process is a definite value. As a result, we use Silhouette

¹There are 27 shared medical features between these institutions, while there are also about 40 private features in each hospital respectively.

Coefficient, Calinski Harabasz Score and Davies Bouldin Score to figure out the appropriate *k*. Algorithm 1 shows the process of the federated clustering method.

Algorithm 1 Federated clustering method

```

server initialize center of k clusters  $c_1, c_2, \dots, c_k$ 
while not convergence do
  for each client  $C_m$  in parallel do
    receive  $c_1, c_2, \dots, c_k$  from server
    for each sample in  $C_m$  do
      compute Euclidean distance of each cluster's center
      send result to server
    end for
  end forserver update center of k clusters
end while

```

5 EXPERIMENTS

We connect the isolated COVID-19 institutions (HM Hospital in Spain [15] and Tongji Hospital in China [36]) and the cardiology institution (PhysioNet in America) to perform the length of stay prediction and the sepsis detection, respectively. Our proposed framework MTFed shows superior performance compared to other federated methods and transfer-learning-based methods.

5.1 Medical Institution Collaborators

- **COVID-19 Collaborator: Tongji Hospital (TJH), China.** The medical information of all patients collected between 10 January and 18 February 2020 at Tongji Hospital, China. The average age of the patients was 58.83 years, and 59.7% were male. Of the 375 cases included in the subsequent analysis, 201 recovered from COVID-19 and were discharged from the hospital, while 174 died.

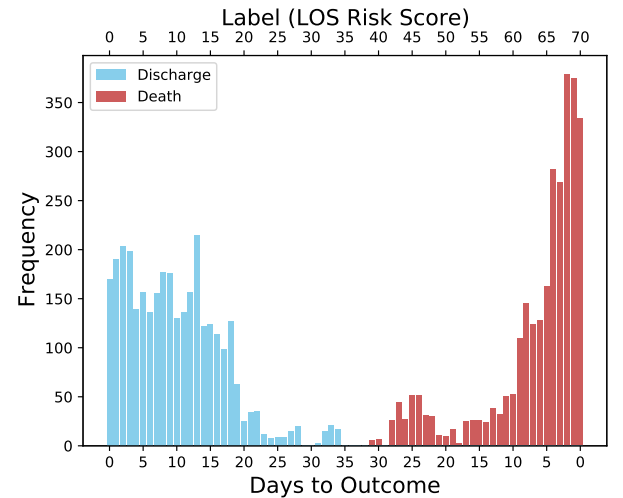


Figure 1: Days to outcome of patients' records in Tongji Hospital, China. All patients discharged or died within 35 days.

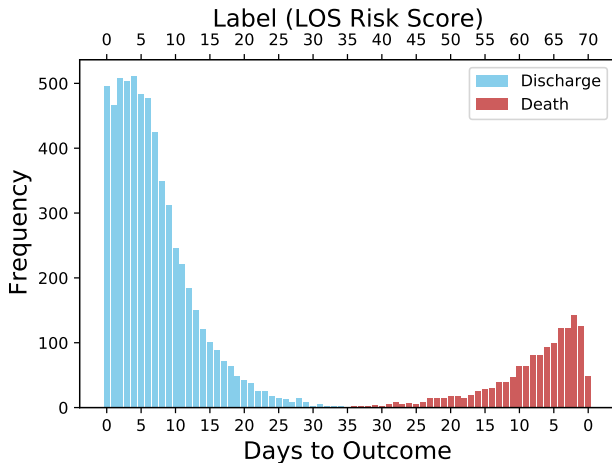
Table 1: Prediction Performance on COVID-19 collaborators. The corner mark indicates the number of collaborators.

Methods	Federated	LOS Prediction Performance on TJH		LOS Prediction Performance on HMH	
		MSE	MAE	MSE	MAE
GRU	×	136.5404(66.4238)	7.6595(2.0569)	392.8527(46.3623)	12.4133(0.8761)
TimeNet [12]	✓	214.1021(30.7063)	11.4234(0.9525)	447.1890(152.3652)	14.8670(2.5631)
MLP-F [34]	✓	186.1036(46.1807)	9.8169(1.2887)	400.9878(13.8591)	13.4772(0.5369)
T-LSTM [1]	×	150.9333 (72.4362)	8.0356 (2.0617)	425.8102(102.9429)	13.5431(2.2985)
MTFed _{col-0}	×	123.0640(45.9865)	7.4360(1.2296)	369.8694(32.4589)	12.1767(1.0955)
MTFed _{col-1}	✓	109.6348(37.9988)	7.3866(1.4728)	358.0318(38.3597)	11.8875(0.7675)
MTFed _{col-2}	✓	101.8956(32.0468)	6.9787(1.0627)	355.3907(45.8098)	11.8420(1.2188)

Table 2: Prediction Performance on cardiology sepsis collaborator. The corner mark indicates the number of collaborators.

Methods	Federated	AUPRC	AUROC	min(Se, P+)
GRU	×	0.6036 (0.0088)	0.9049 (0.0057)	0.5766 (0.0143)
TimeNet [12]	✓	0.5398 (0.0097)	0.8444 (0.0089)	0.5038 (0.0189)
MLP-F [34]	✓	0.3075 (0.0164)	0.8146 (0.0098)	0.3600 (0.0172)
T-LSTM [12]	×	0.6503 (0.0249)	0.9180 (0.0071)	0.6287 (0.0224)
MTFed _{col-0}	×	0.7053 (0.0069)	0.9319 (0.0048)	0.6537 (0.0077)
MTFed _{col-1}	✓	0.7190(0.0097)	0.9352(0.0057)	0.6598(0.0130)

- **COVID-19 Collaborator: HM Hospitals (HMH), Spain.** HMH is released by HM Hospitals containing 2,310 anonymous patients who were diagnosed with COVID-19 or to be confirmed. These data collect various interactions during the treatment of COVID-19, including detailed information about the diagnosis, treatment, admission, steps through the ICU, and discharge or death. We selected the patients who have at least one record of lab tests and indicators with a missing rate of higher than 10% are dropped. After screening, there are 1,891 patients, and 303 patients died. The distribution of length of stay in HM Hospital is shown in Figure 2.

**Figure 2: Days to outcome of patients' records in HM Hospital, Spain. Most patients discharged or died within 35 days.**

- **Cardiology Sepsis Collaborator: PhysioNet, America (PHY).**

² The PhysioNet [30] collaborator consists of three geographically distinct U.S. hospital systems with three different electronic medical record systems. These data were collected over the past decade with approval from the appropriate Institutional Review Boards. The cleaned dataset consists of 40,336 patients and consists of hourly vital sign summaries, lab values, and static patient descriptions. In particular, the data contained 40 clinical variables: 8 vital sign variables (e.g., heart rate, systolic blood pressure), 26 laboratory variables (e.g., Chloride, Glucose), and 6 demographic variables. **Problem: Sepsis Detection for Cardiology Collaborator** can be formulated as a binary classification problem and labeled by sepsis-3 clinical criteria.

5.1.1 Baseline Approaches. We introduce several deep-learning-based models as our baseline approaches without additional labeled data or external ontology resources.

- GRU [6] is the basic Gated Recurrent Unit network.
- T-LSTM (SIGKDD) [1] handles irregular time intervals by time decay mechanism. We modify it into a supervised learning model.
- TimeNet (IJCAI) [12] maps variable-length clinical time series to fixed-dimensional feature vectors separately, and acts as an off-the-shelf feature extractor. It is pre-trained on the UCR time series Repository.
- MLP-F [34] employs federated learning to predict the mortality for COVID-19 patients using Multilayer Perceptron.
- MTFed_{col-num} is the proposed MTFed with several collaborators. The institution trains the model only on its dataset when $num = 0$.

²<https://physionet.org/>

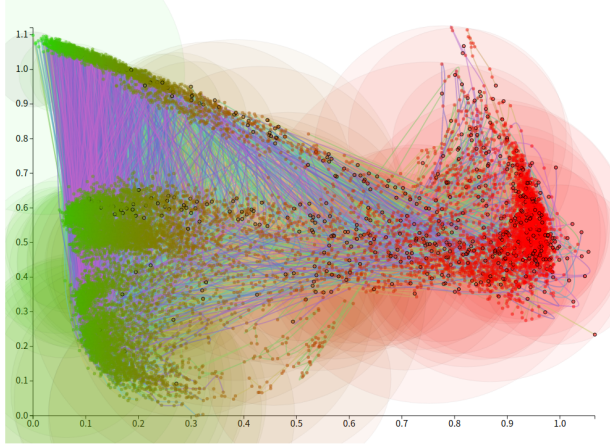


Figure 3: Visualization of COVID-19 patients' health trajectory in TJH and HMH.

5.2 Quantitative Analysis

As is shown in Table 1, the federated method MTFed consistently outperforms non-federated baselines, demonstrating the ability of our model to learn robust representation. Especially on TJH, MTFed decreases the MSE to 101.90 and MAE to 6.98, which is much lower than the comparative baselines. Besides, our model achieves a relatively lower MSE of 14.48 on HMH. It is evident that compared to federated trained models with two collaborators, the models trained with three collaborators are superior in performance. With the help of MTFed, smaller communities and rural hospitals would enjoy access to expert-level AI algorithms.

Table 2 reports the performance of our model on PhysioNet. Though the sample size of PhysioNet is the biggest, its task and even type of the disease are different from the other, there is still 2% promotion on AUPRC via MTFed, which confirms the robustness of our model to adapt to different situations.

5.3 Qualitative Results

5.3.1 Demonstration of Federated Clustering Results. Additionally, after clustering on TJH and HMH, we use PCA (principal components analysis) to reduce the dimensionality of patients' representation on each timestep and visualize them on a 2-dimension plane, as shown in Figure 3, where each circle represents a cluster and each visit is denoted as a point. The line with several points represents the health status trajectory of a patient and the color reflects the predicted length-of-stay of a patient at that time. Red means high risk and green means low risk, or in other words, a relatively more healthy status. Black hollow circle means death. The statistics for clusters (e.g. mortality rate, gender ratio) will help the doctors to analyze the health conditions of the groups, predict their future disease progression, and extract medical knowledge.

6 CONCLUSION

In this work, we propose a federated learning model MTFed, which encourages different hospitals, healthcare institutions and research centers to collaborate on building a model that could benefit them

all. MTFed allows every participant keeps control of its own clinical data without needing to directly share any local sensitive raw data. MTFed is enabled by a multi-channel architecture which can embed each clinical feature separately in clinical representation learning to applied across institutions with different feature sets. The experiments on real-world collaborators show that MTFed can improve the performance of health evaluation effectively for all collaborators. This performance improvement is especially useful under data insufficiency settings, especially at the early stage of the emerging pandemic.

REFERENCES

- [1] Inci M Baytas, Cao Xiao, Xi Zhang, Fei Wang, Anil K Jain, and Jiayu Zhou. 2017. Patient subtyping via time-aware LSTM networks. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 65–74.
- [2] Edward Choi, Mohammad Taha Bahadori, Andy Schuetz, Walter F Stewart, and Jimeng Sun. 2015. Doctor AI: Predicting Clinical Events via Recurrent Neural Networks. (2015).
- [3] Edward Choi, Mohammad Taha Bahadori, Le Song, Walter F Stewart, and Jimeng Sun. 2017. GRAM: graph-based attention model for healthcare representation learning. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 787–795.
- [4] Edward Choi, Mohammad Taha Bahadori, Jimeng Sun, Joshua Kulas, Andy Schuetz, and Walter Stewart. 2016. Retain: An interpretable predictive model for healthcare using reverse time attention mechanism. In *Advances in Neural Information Processing Systems*. 3504–3512.
- [5] Edward Choi, Cao Xiao, Walter Stewart, and Jimeng Sun. 2018. Mime: Multilevel medical embedding of electronic health records for predictive healthcare. In *Advances in Neural Information Processing Systems*. 4547–4557.
- [6] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555* (2014).
- [7] Amir Emami, Fatemeh Javanmardi, Neda Pirbonyeh, and Ali Akbari. 2020. Prevalence of underlying diseases in hospitalized patients with COVID-19: a systematic review and meta-analysis. *Archives of academic emergency medicine* 8, 1 (2020).
- [8] Saba Emrani, Anya McGuirk, and Wei Xiao. 2017. Prognosis and Diagnosis of Parkinson's Disease Using Multi-Task Learning. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 1457–1466.
- [9] Cristóbal Esteban, Oliver Staack, Stephan Baier, Yinchong Yang, and Volker Tresp. 2016. Predicting clinical events by combining static and dynamic information using recurrent neural networks. In *Healthcare Informatics (ICHI), 2016 IEEE International Conference on*. Ieee, 93–101.
- [10] Leiwen Fu, Bingyi Wang, Tanwei Yuan, Xiaoting Chen, Yunlong Ao, Tom Fitzpatrick, Peiyang Li, Yiguo Zhou, Yifan Lin, Qibin Duan, et al. 2020. Clinical characteristics of coronavirus disease 2019 (COVID-19) in China: a systematic review and meta-analysis. *Journal of Infection* (2020).
- [11] Priyanka Gupta, Pankaj Malhotra, Lovekesh Vig, and Gautam Shroff. 2018. Transfer Learning for Clinical Time Series Analysis using Recurrent Neural Networks. *arXiv preprint arXiv:1807.01705* (2018).
- [12] Priyanka Gupta, Pankaj Malhotra, Lovekesh Vig, and Gautam Shroff. 2018. Using Features from Pre-trained TimeNet for Clinical Predictions. In *The 3rd International Workshop on Knowledge Discovery in Healthcare Data at IJCAI*.
- [13] Hrayr Harutyunyan, Hrant Khachatryan, David C Kale, and Aram Galstyan. 2017. Multitask learning and benchmarking with clinical time series data. *arXiv preprint arXiv:1703.07771* (2017).
- [14] Jay Heo, Hae Beom Lee, Saehoon Kim, Juho Lee, Kwang Joon Kim, Eunho Yang, and Sung Ju Hwang. 2018. Uncertainty-aware attention for reliable interpretation and prediction. In *Advances in Neural Information Processing Systems*. 909–918.
- [15] HM hospitales. [n.d.]. COVID DATA SAVE LIVES. <https://www.hmhosptales.com/>. Accessed: 2020-10-20.
- [16] Chaolin Huang, Yeming Wang, Xingwang Li, Lili Ren, Jianping Zhao, Yi Hu, Li Zhang, Guohui Fan, Jiuyang Xu, Xiaoying Gu, et al. 2020. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *The lancet* 395, 10223 (2020), 497–506.
- [17] Xiaowei Huang, Daniel Kroening, Wenjie Ruan, James Sharp, Youcheng Sun, Emese Thamo, Min Wu, and Xinpeng Yi. 2020. A survey of safety and trustworthiness of deep neural networks: Verification, testing, adversarial attack and defence, and interpretability. *Computer Science Review* 37 (2020), 100270.
- [18] Yunpeng Ji, Zhongren Ma, Maikel P Peppelenbosch, and Qiuwei Pan. 2020. Potential association between COVID-19 mortality and health-care resource availability. *The Lancet Global Health* 8, 4 (2020), e480.

- [19] Changhee Lee, William R Zame, Jinsung Yoon, and Mihaela van der Schaar. 2018. Deephit: A deep learning approach to survival analysis with competing risks. AAAI.
- [20] Wonsung Lee, Sungrae Park, Weonyoung Joo, and Il-Chul Moon. 2018. Diagnosis Prediction via Medical Context Attention Networks Using Deep Generative Modeling. In *2018 IEEE International Conference on Data Mining (ICDM)*. IEEE, 1104–1109.
- [21] Boyi Liu, Bingjie Yan, Yize Zhou, Yifan Yang, and Yixian Zhang. 2020. Experiments of federated learning for covid-19 chest x-ray images. *arXiv preprint arXiv:2007.05592* (2020).
- [22] Fenglong Ma, Quanzeng You, Houping Xiao, Radha Chitta, Jing Zhou, and Jing Gao. 2018. Kame: Knowledge-based attention model for diagnosis prediction in healthcare. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*. ACM, 743–752.
- [23] Liantao Ma, Chaohe Zhang, Yasha Wang, Wenjie Ruan, Jiangtao Wang, Wen Tang, Xinyu Ma, Xin Gao, and Junyi Gao. 2020. Concare: Personalized clinical feature embedding via capturing the healthcare context. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 833–840.
- [24] Tengfei Ma, Cao Xiao, and Fei Wang. 2018. Health-ATM: A Deep Architecture for Multifaceted Patient Health Record Representation and Risk Prediction. In *Proceedings of the 2018 SIAM International Conference on Data Mining*. SIAM, 261–269.
- [25] 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.
- [26] US Department of Health, Human Services, et al. 2017. Pandemic influenza plan: 2017 Update. URL <https://www.cdc.gov/flu/pandemic-resources/pdf/pan-flu-report-2017v2.pdf> (2017).
- [27] Trang Pham, Truyen Tran, Dinh Phung, and Svetha Venkatesh. 2016. Deepcare: A deep dynamic memory model for predictive medicine. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer, 30–41.
- [28] Sanjay Purushotham, Chuizheng Meng, Zhengping Che, and Yan Liu. 2017. Benchmark of Deep Learning Models on Large Healthcare MIMIC Datasets. *arXiv: Learning* (2017).
- [29] Zhi Qiao, Shiwan Zhao, Cao Xiao, Xiang Li, Yong Qin, and Fei Wang. 2018. Pairwise-Ranking based Collaborative Recurrent Neural Networks for Clinical Event Prediction.. In *IJCAI*. 3520–3526.
- [30] Matthew A Reyna, Chris Josef, Salman Seyed, Russell Jeter, Supreeth P Shashikumar, M Brandon Westover, Ashish Sharma, Shamim Nemati, and Gari D Clifford. 2019. Early Prediction of Sepsis from Clinical Data: the PhysioNet/Computing in Cardiology Challenge 2019. (2019).
- [31] Micah J. Sheller, Brandon Edwards, G. Anthony Reina, Jason Martin, and Spyridon Bakas. 2020. Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data. *entific Reports* 10, 1 (2020).
- [32] 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. (2018).
- [33] Harini Suresh, Jen J Gong, and John Guttag. 2018. Learning Tasks for Multi-task Learning: Heterogenous Patient Populations in the ICU. *arXiv preprint arXiv:1806.02878* (2018).
- [34] Akhil Vaid, Suraj K Jaladanki, Jie Xu, Shelly Teng, Arvind Kumar, and Samuel Lee. [n.d.]. Federated Learning of Electronic Health Records Improves Mortality Prediction in Patients. *Ethnicity* 52, 77.6 ([n. d.]), 0–001.
- [35] Xinhui Wang, Xuexian Fang, Zhaoxian Cai, Xiaotian Wu, Xiaotong Gao, Junxia Min, Fudi Wang, et al. 2020. Comorbid Chronic Diseases and Acute Organ Injuries Are Strongly Correlated with Disease Severity and Mortality among COVID-19 Patients: A Systemic Review and Meta-Analysis. *Research* 2020 (2020), 2402961.
- [36] Li Yan, Hai-Tao Zhang, Jorge Goncalves, Yang Xiao, Maolin Wang, Yuqi Guo, Chuan Sun, Xiuchuan Tang, Liang Jing, Mingyang Zhang, et al. 2020. An interpretable mortality prediction model for COVID-19 patients. *Nature Machine Intelligence* (2020), 1–6.
- [37] Ye Yuan, Guangxu Xun, Qiuling Suo, Kebin Jia, and Aidong Zhang. 2017. Wave2vec: Learning deep representations for biosignals. In *Data Mining (ICDM), 2017 IEEE International Conference on*. IEEE, 1159–1164.
- [38] W. Zhang, Tao Zhou, Qinghua Lu, X. Wang, Chunsheng Zhu, Haoyun Sun, Zhipeng Wang, Sin Kit Lo, and F.-Y. Wang. 2020. Dynamic Fusion based Federated Learning for COVID-19 Detection. *ArXiv abs/2009.10401* (2020).
- [39] Kaiping Zheng, Jinyang Gao, Kee Yuan Ngiam, Beng Chin Ooi, and Wei Luen James Yip. 2017. Resolving the bias in electronic medical records. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2171–2180.