JAMIA Open, 2024, **7(4)**, ooae110
https://doi.org/10.1093/jamiaopen/ooae110
Research and Applications

Research and Applications

Federated learning of medical concepts embedding using BEHRT

Ofir Ben Shoham, BSc[1] and Nadav Rappoport, PhD[1,*]

[1]Department of Software and Information Systems Engineering, Ben-Gurion University of the Negev, Be'er Sheva, Israel

*Corresponding author: Nadav Rappoport, PhD, Department of Software and Information Systems Engineering, Ben-Gurion University of the Negev, Be'er Sheva, Israel (nadavrap@bgu.ac.il)

Abstract

Objectives: Electronic health record data is often considered sensitive medical information. Therefore, the EHR data from different medical centers often cannot be shared, making it difficult to create prediction models using multicenter EHR data, which is essential for such models' robustness and generalizability. Federated learning (FL) is an algorithmic approach that allows learning a shared model using data in multiple locations without the need to store all data in a single central place. Our study aims to evaluate an FL approach using the BEHRT model for predictive tasks on EHR data, focusing on next visit prediction.

Materials and Methods: We propose an FL approach for learning medical concepts embedding. This pretrained model can be used for fine-tuning for specific downstream tasks. Our approach is based on an embedding model like BEHRT, a deep neural sequence transduction model for EHR. We train using FL, both the masked language modeling (MLM) and the next visit downstream model.

Results: We demonstrate our approach on the MIMIC-IV dataset. We compare the performance of a model trained with FL to one trained on centralized data, observing a difference in average precision ranging from 0% to 3% (absolute), depending on the length of the patients' visit history. Moreover, our approach improves average precision by 4%-10% (absolute) compared to local models. In addition, we show the importance of the usage of pretrained MLM for the next visit diagnoses prediction task.

Discussion and Conclusion: We find that our FL approach reaches very close to the performance of a centralized model, and it outperforms local models in terms of average precision. We also show that pretrained MLM improves the model's average precision performance in the next visit diagnoses prediction task, compared to an MLM without pretraining.

Lay Summary

Electronic health records (EHRs) contain sensitive medical information that is crucial for improving patient care. However, sharing this data between different medical centers can be challenging due to privacy concerns. Our study explores a solution using a technique called federated learning (FL), which allows multiple institutions to collaborate on building predictive models without needing to share their data. We focused on predicting a patient's next visit based on their health history, using a specialized model called BEHRT, which processes EHR data effectively. By applying, we trained this model across different centers while keeping the data secure at each location. Our findings show that the FL approach can achieve results comparable to traditional methods that use centralized data. Additionally, we discovered that using a pretrained method for understanding medical language significantly improves the model's predictions. Overall, our work highlights how FL can enhance healthcare analytics while protecting patient privacy, paving the way for more robust and generalizable health predictions across multiple institutions.

Key words: federated learning; NLP; electronic health records (EHRs); machine learning; prediction model; BERT.

Introduction

Electronic health records (EHR) is a collection of pieces of information documenting a patient's medical history (for example, patient's drug prescriptions and admissions to medical centers). The medical records stored in medical centers contain critical medical information about the treatment protocol and its results.[1]

Multicenter studies have the potential to enhance models' ability to capture and adapt to heterogeneity, leading to an improvement in their generalizability. Furthermore, collecting data from multiple sources results in a larger dataset for training prediction models, which reduces the expected generalization error and increases the robustness of the model.[2] In addition, rare conditions may not be represented well enough in a single dataset. However, incorporating data from multiple sources can enhance the representation of these conditions and increase their significance in training machine learning models.[3]

Electronic health records contain sensitive medical information, which can make it challenging to share among healthcare providers. Federated learning (FL) is an algorithmic approach that trains a single model based on several databases stored in separate locations (clients) without consolidating the information in 1 central location.[4] This approach makes it possible to train a shared global machine learning model with the help of a central server without sharing the observations outside their authorized location. In particular, FL is suitable for training a computational model based on information sources from separate medical centers (multicenter study) while maintaining the privacy of data, patients, and medical centers.[5]

Received: July 26, 2024; **Revised:** September 26, 2024; **Editorial Decision:** October 3, 2024; **Accepted:** October 6, 2024

An example of a prediction task based on EHR data is the prediction of future diagnoses, also called next visit prediction. In this task, we want to train a model that can predict the diagnoses of a patient that will be diagnosed with in his next visit based on current and previous clinical data.

BEHRT[6] is a deep neural sequence transduction model based on BERT[7] architecture for EHR. The input for this model is a sequence constructed with words representing diagnoses, sentences representing each visit, and a document representing a patient's complete medical history. In their work, Yikuan et al. first trained an MLM and then used it as a pretrained model and fine-tuned it for the next visit prediction task. Afterward, Yikuan et al. demonstrated their approach on the CPRD dataset[8] that contains medical records from general practitioners. BEHRT demonstrates an enhancement of 8.0%-13.2% (in terms of average precision scores for various tasks) compared to the state-of-the-art deep EHR models like RETAIN[9] and Deepr[10] models.[6] The BEHRT architecture is designed to easily incorporate multiple heterogeneous medical concepts, including diagnoses, measurements, and more. Furthermore, BEHRT's patient representation can be used as a pretrained model for downstream tasks.[6]

However, the training of the BEHRT model, as discussed in Li et al.,[6] is limited by centralizing all data, which prevents the BEHRT model from handling multicenter data. We proposed FL training to medical concepts embedding using BEHRT. Our approach utilized FL training to enhance the robustness and generalizability of the BEHRT model. Our approach used FL training for the pretrained MLM phase and also for the next visit prediction task. Our approach is applicable to any dataset containing clinical data per patient and is suitable for multicenter studies that require an FL algorithm to ensure EHR data privacy. In this work, we explored the use of FL to train BEHRT, aiming to preserve privacy with minimal impact on performance.

We demonstrated our approach using the MIMIC-IV dataset[11] for the next visit prediction task. Our FL approach improves average precision by 4-10 absolute percents compared to local models, and achieves very close average precision performance to centralized models, while maintaining data privacy and scalability for multicenter studies.

Related work

Our work relies on a representation model of medical concepts. In recent decades, word2vec methods have gained popularity not only in classical NLP but also in Precision Medicine.[12] For example, Phe2Vec[13] generates patient embeddings by representing a patient's medical history through medical concepts such as diagnoses, procedures, lab tests, and medications. These medical concepts are grouped into intervals of days, with each interval treated as a sentence and each medical concept within that interval represented as a word. The resulting sequence of words allows for the application of word embedding techniques, including methods like GloVe,[14] FastText,[15] or BERT,[7] which is based on the transformer architecture.[16] BRLTM utilized transformers (MLM and afterward fine-tuned the pretrained model) to predict depression.[17] Other representation models include SapBERT, which captures fine-grained semantic relationships in the biomedical domain,[18] and BERGAMOT, which combines pretrained language models with graph neural networks to capture both interconcept and interconcept interactions from the multilingual UMLS graph.[19] Large language models (LLMs) can also be used for representation learning, as demonstrated by Ronzano and Nanavati.[20] However, utilizing LLMs for this purpose requires significant computational resources due to the high number of parameters in these models.

Another common transformer-based model is Med-BERT.[21] Med-BERT trains BERT model using the MLM task, then trains the model for length-of-stay (LOS) task. Med-BERT demonstrated an improvement on 2 downstream tasks compared to GRU and RETAIN,[9] but there is no comparison to BEHRT. The main differences between Med-BERT and BEHRT is that Med-BERT was trained also on the LOS task and has more training samples compared to BEHRT. However, Med-BERT has a ranking for each event, and the ranking of the importance of each event has not been studied enough.[21] In addition, Rasmy et al. did not include the time between different visits, unlike BEHRT. Therefore, we chose to illustrate our approach using BEHRT.

We chose to evaluate our approach in a manner similar to how BEHRT was evaluated, focusing on next visit diagnoses prediction. Previous studies have developed models for predicting diagnoses at the next patient admission based on medical history. Gao et al. proposed co-attention memory networks for diagnosis prediction, which enhances RNNs with a memory network to improve representation capacity.[22] Ma et al. introduced a diagnosis prediction framework that integrates diagnosis code embeddings with a predictive model.[23] While there are additional studies that address predicting diagnoses for subsequent admissions, our focus was on evaluating BEHRT's performance for this task. Specifically, we aimed to assess how well BEHRT can be trained to maintain privacy (using FL) while achieving performance comparable to other methods in predicting next visit diagnoses.

Med-BERT used multicenter data, but all the data located in one central location, which limits the available data due to concerns over infrastructures, regulations, privacy, and data standardization present a challenge to data sharing across healthcare institutions.[2] Multicenter EHR data enables the larger and varied data for model training which is essential in order to improve model generalizability and robustness.[2] There are federated algorithms to overcome this limitation such as the literature.[4,24]

In healthcare, previous studies have explored the use of FL. Boughorbel et al. applied FL for early birth prediction, where the model's contribution to the aggregation decreases if its confidence is low.[25] Grama et al. utilized FL with a neural network for disease prediction.[26] Additionally, FedEHR applied FL for heart disease prediction using EHR data sourced from Internet of Things devices.[27]

While these works highlight the potential of FL for healthcare tasks, several general FL algorithms have been developed and evaluated across various applications. For example, FedBERT[28] employed FedAvg[4] for BERT pretraining but does not focus on clinical aspects, such as medical concepts. Dang et al.[2] compared multiple FL algorithms such as FedAvg, FedAvgM, FedProx, FedAdam, and FedAdagrad. Among the FL algorithms, the FedAvg and FedAvgM algorithms achieved slightly better results than FedProx, FedAdam, and FedAdagrad.[2] In this work, we adopted the FedAvg algorithm for FL, as it is a widely adopted and commonly used

method. We used the FedAvg FL algorithm for the MLM and next visit prediction tasks. We used transformer-based modeling according to the model architecture of BEHRT.[6]

Methods

Initially, we retrieved the data from the raw source MIMIC-IV database 4.2. Next, we simulate a federated data scenario by dividing the data into multiple centers. Each patient was assigned to a single center according to the center where it had the longest stay in. Afterward, we employed FL training for MLM. Lastly, we utilized the MLM pretrained model for FL of the next visit prediction task. We selected this task for evaluation to enable a comparison of our approach with the task defined in BEHRT, as well as the metrics for comparison. An overview of the stages of this study is illustrated in the Supplementary Material (Figure S1).

Next visit problem definition

We adhered to the problem definition for the next visit prediction task as presented in BEHRT.[6] Let P denote the set of patients and let each patient p have medical data consisting of n visits: $V_p = V_{1,p}, V_{2,p}, \ldots, V_{n,p}$. For a given visit i of patient p, $V_{i,p}$ represents the set of diagnoses assigned to patient p at visit i. Specifically, $V_{i,p} = d_{1,i,p}, d_{2,i,p}, \ldots, d_{m,i,p}$, where m is the number of diagnoses assigned to patient p at visit number i. In the next visit prediction task, we choose a random j visit number, assuming we have the medical data until V_j, we need to predict the diagnoses $d_{1,j+1,p}, d_{2,j+1,p}, \ldots, d_{m,j+1,p}$ for visit number $j+1$ based on $V_{1,p}, V_{2,p}, \ldots, V_{j,p}$.

Data

MIMIC-IV[11] is a comprehensive healthcare dataset that was utilized to demonstrate the usability of our suggested approach. The complete preprocessing of this research is detailed in the Supplementary Material and illustrated in Figure S2. Each observation is a sequence that represents the medical history of a single patient, which includes his diagnoses, age, and year of diagnosis. The data for each patient is actually composed of multiple visits, ordered by admission start time, which is important for the next visit prediction task. In addition, similar to BEHRT, there is no prescribed order for the multiple diagnoses within a visit.[6] The order is for the visits, but within a visit there is no order for the diagnoses.

Multicenters' split

To demonstrate the need for FL for the next visit prediction task, we simulated a multicenter scenario by splitting our data by patient. To simulate a real-world biased variety between medical centers, we did not split the patients randomly but clinically-driven. Each patient was assigned to a single care unit according to the unit with the longest stay. Each patient was assigned to only 1 center. Length of stay was taken from the MIMIC-IV transfers table in Hosp module.[11] The number of centers is not predefined, but rather determined by the number of different care units. After splitting the patients into centers, we obtained a total of 39 centers.

Baseline approaches

In order to compare our FL approach, we trained a centralized model. In the centralized training, the 2 learning phases of MLM and next visit prediction were trained using a single dataset covering all the training samples. In addition, we also compared our approach to local model training. In the local training, no information is shared across clients. As we have 39 centers, we trained each center's model separately using its local data. First of all, we trained MLM for the local data, and then we fine-tuned the MLM using the client's local data for the next visit prediction.

BEHRT

We used the BEHRT[6] model architecture for FL for both the MLM and the next visit prediction downstream task. BEHRT is a deep learning model built upon the BERT architecture.[7] BEHRT consists of MLM that was fine-tuned by adding a classification layer for the next visit prediction task. In the MLM training, the task is to predict the masked disease tokens. The features for the MLM tasks are: diagnoses, patient's age, and the diagnosis year. The embedding layer in BEHRT is constructed through the concatenation of these features. For the next visit task, the features are the same as those for the MLM, but the list of diagnoses is partial and contains the medical information up to the visit for which we want to predict its diagnoses. We used the same features of BEHRT. In the MLM phase, the model learns an embedding of the clinical concepts such as diagnosis, age, position (ie, the relative position of a concept within a visit), and segment (ie, visit). Afterward, the MLM is fine-tuned for next visit prediction by adding a classification layer. In addition, the pretrained BEHRT model (after the MLM phase) can be utilized for tasks beyond next visit prediction through transfer learning.[6]

Our approach

We used the Federated Averaging (FedAvg) algorithm[4] for training BEHRT with FL, motivated by the need to preserve patient privacy and increase the generalizability of the trained model. FedAvg is an FL approach that aggregates locally computed model updates from multiple devices or institutions into a global model.

The following steps outline our proposed approach for implementing FL to train the BEHRT model. First, The server initially shares the BEHRT global model with each client. Subsequently, the selected client trains their local model using their local data as depicted in Figure 1(A). In the second step, the selected clients transmit the weights of their trained local model to the server, without sharing their local and private data. Then, the server updates the global model by aggregating all the updated models by computing a weighted average of each weight according to the client's sample size, as shown in Figure 1(B). This weighted average is essential, as it ensures that the contributions from clients with larger patient samples are appropriately prioritized, addressing the differences in patient numbers among hospitals. Finally, the server disseminates the updated model to all the clients. The iterative process continued for 500 epochs, after which the best model was selected based on the highest average precision achieved on the validation set. We used this FL (FedAvg) algorithm for both the MLM training step and the next visit prediction model training step. First, we apply our FL approach to train the MLM. Then, we fine-tune this MLM for the next visit prediction task by adding a classification layer. At each round

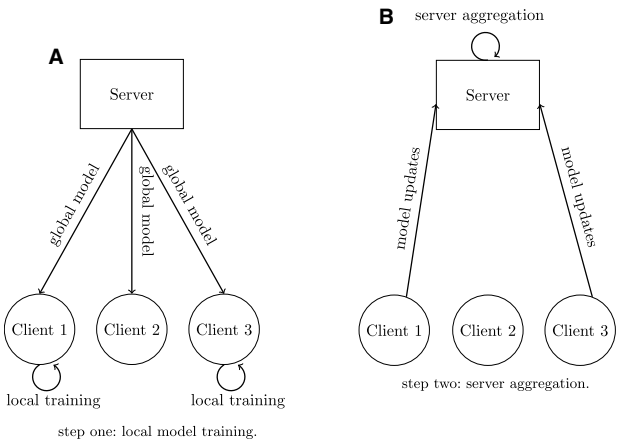

Figure 1. Federated learning algorithm for MLM and next visit prediction. (A) In the first step, the server sends the global model to all clients, and each selected client trains the local model. (B) In the second step, the server gets the trained weights from the selected clients, aggregates the weights, and updates the global model. Abbreviation: MLM, masked language modeling.

of training, we selected only a fraction of 10% from the clients to train on their local data. We did this for efficiency, as McMahan et al.[4] showed that there is a point of diminishing returns when adding more clients.

Experiments

We performed multiple experiments to compare our proposed FL approach to a model trained with all the data in a central place. We repeated the training of the next visit prediction while varying the random seed in order to calculate the CIs. More details regarding the implementation details can be found in the Supplementary Material.

Federated vs centralized learning

In this experiment, we compared our proposed approach (FL training) to a model trained with centralized data. For the FL training, the two phases (FL MLM and FL next visit prediction) were trained using the federated data. We trained a single MLM and multiple next-visit models where in each model, we subset the data to patients having at least 1, 3, 5, or 15 visits. Our results showed that our proposed FL model achieved similar average precision to the centralized model for minimum visits of 3, 5, and 15. For a minimum visit of 1, the centralized model outperformed our model by an absolute value of 3% (Figure 2).

Federated vs local client-independent learning

This experiment simulates a scenario where no data can be shared due to privacy and security concerns, making local model training a common scenario in such cases. Each local model was trained with its own local data, which varied in size and clinical conditions. To aggregate the performances of the local models, we used weighted averages based on their average precision and the number of examples (patients) in the local train data. Figure 2 shows the average precision results of local training compared to FL training and centralized training for 4 minimum visit thresholds. Our FL approach outperformed local training for minimum visits of 1, 3, 5, and 15 by an average precision of absolute 4%, 8%, 8%, and 10%, respectively. Overall, our proposed FL training model achieved 4%-10% absolute higher average precision than local training models.

Pretrained MLM

In the next step, we took the pretrained MLMs and fine-tuned them for the prediction task. In this experiment, we conducted an ablation study to evaluate the importance of pretrained MLM. Specifically, we compared the performance of FL next visit prediction using different pretrained MLMs. We evaluated 2 centralized MLMs: the first was an MLM with at least 1 visit (trained on all patients), and the second was an MLM trained on patients who had at least 3 visits. Additionally, we evaluated 2 more FL MLMs. The first FL MLM is trained with patients who had at least 1 visit, and the second is for patients with at least 3 visits. Finally, we compared the performance of all these pretrained models to the performance of the model without pretrained MLM. Figure 3 shows the average precision comparison of FL next visit prediction based on the pretrained MLMs. This figure shows that for minimum visits of 3 and 5, the pretrained MLM improves the average precision for FL next visit prediction by 1%-1.2% absolute compared to without pretrained MLM. Moreover, the difference in average precision between the centralized MLM and FL MLM was negligible. These findings indicate that FL MLM can achieve similar performance without having all the data centralized in one place.

Discussion

In this article, we present an FL approach for BEHRT. We trained the MLM and the next visit prediction task using the FedAvg algorithm.[4] Our approach is general and well suited for multicenter studies that require an FL model to ensure the privacy of EHR data. We show that our approach of FL of embedding clinical concepts can meet the performance of a model trained on centralized data, and it outperforms model trained locally with no information sharing. We demonstrate the effectiveness of our approach by simulating the MIMIC-IV dataset as a multicenter study, training an FL MLM and next visit prediction models. We compare the performance of our FL approach to both a centralized model and local models (which are commonly used due to data privacy concerns).

In the first experiment, we compare the average precision of FL training, centralized training, and local training for different minimum visit thresholds. Our FL approach achieved average precision results that were comparable to the centralized baseline approach. For minimum visit thresholds of 3, 5, and 15, the differences in average precision were negligible. These results demonstrate that our approach can achieve similar performance to centralized training while preserving EHR data privacy. The reason for lower performance for a minimum visit threshold of 1 is not clear enough. One possible reason is that the set of diagnoses of the patients in this dataset are more diverse, which could make it more difficult for the FL model to generalize well across all clients. In contrast, for minimum visits threshold of 3 and above the sample size is smaller and the set of possible diagnoses and concept to learn their embedding is smaller.

We compare our approach to local models, where each center trains with its local data. We find that the difference in performance between local training and our FL approach increased as the minimum visit threshold increased from 1 to 3 and from 5 to 15 (Figure 2). A possible reason for the

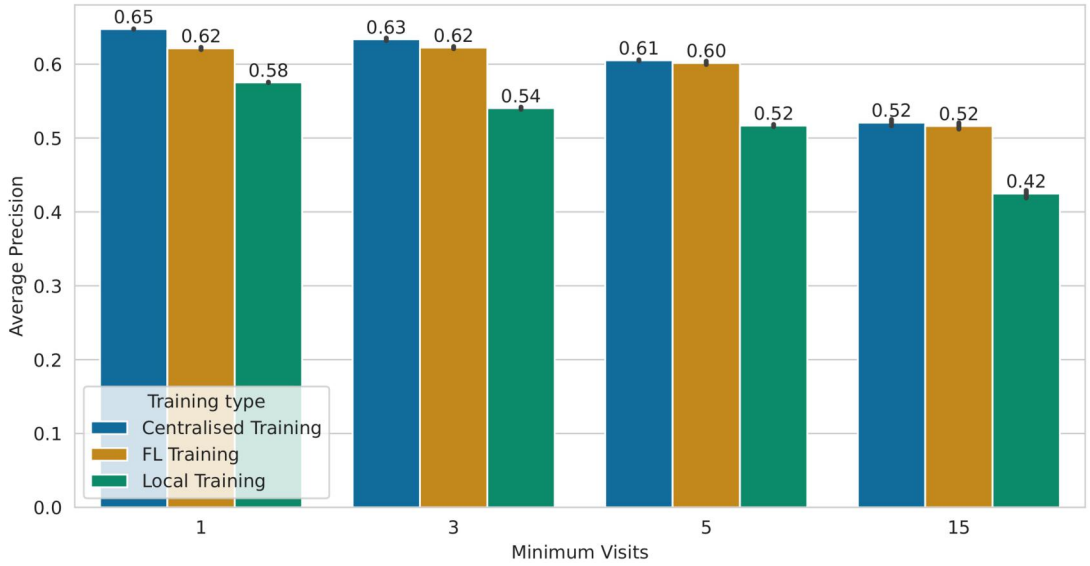

Figure 2. The average precision of each training method was evaluated for the next visit prediction task. The centralized model is referred to as centralized training. Our proposed approach is FL training, and local training involves training local models in our multicenter study. We evaluated the average precision of the models at 4 minimum visit thresholds. The average precision value appears at the top of each bar plot, and also the 95% CI based on a random seed. Abbreviation: FL, federated learning.

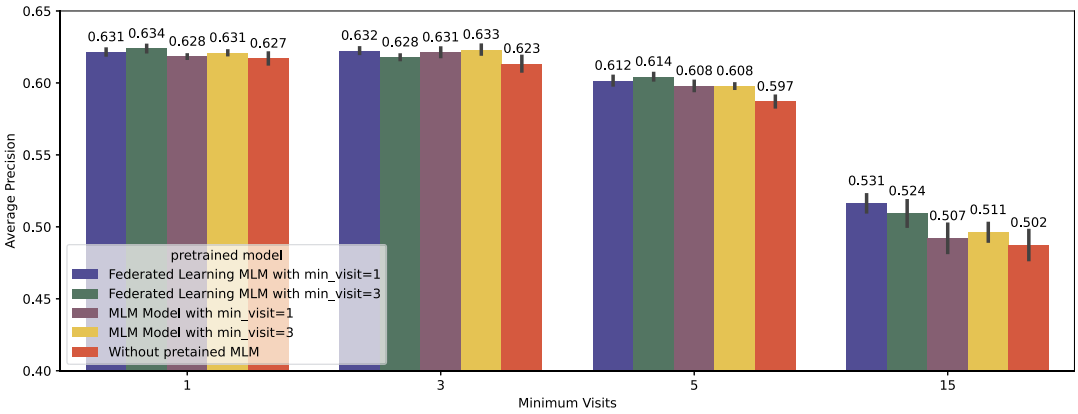

Figure 3. The importance of the pretrained MLM for the FL next visit prediction task. We compared the performance of 5 MLM configurations: blue and green with pretrained MLMs and fine-tuned with FL (for patients with minimum visits of 1 or 3, respectively); purple and yellow with centralized MLM training (also for patients with minimum visits of 1 or 3, respectively); and red without pretrained MLMs. We included the 95% CI on each bar plot. Abbreviations: FL, federated learning; MLM, masked language modeling.

decrease in performance of the local models when increasing the threshold of minimum visits is because local models have less data, making it challenging to learn a local model with good performance. In contrast, the difference between our FL approach and local models is much more significant when there is less data in each center, because the FL approach deals with this by learning a common model, while the local model will have less robustness when it has few examples. In addition, the average precision of the next visit models is lower when the minimum number of visits increases. We believe this is because the number of samples decreases as the minimum visit threshold increases.

In our second experiment, we investigated the impact of using different fine-tuned MLMs for predicting the next medical visit with FL. We found that the performance of centralized MLMs and federated MLMs was similar, but both outperformed the models without pretrained MLMs (Figure 3). These results demonstrate that pretraining the MLMs can significantly improve the average precision of the next visit prediction models. Furthermore, we observed that the performance gap between the pretrained MLMs and the models without pretraining increased as the minimum number of visits per patient increased. This is may be because the pretrained MLMs are particularly valuable in low-data scenarios, where the pretrained MLMs can help to improve the generalization and robustness of the models. Moreover, it can be seen that FL MLM has better performance than a centralized MLM as a pretrained MLM for fine-tuning for the FL next visit prediction (comparing the blue and green bars to purple and yellow bars in Figure 3).

In-depth analysis

Relationship between the minimum number visit threshold and sample size

We observed a decrease in the performance of the next visit models as the minimum number of visits increased.

Table 1. Average precision (AP) and AUC-ROC with CIs for the same number of training data for each minimum visit.

Min visit	Original sample size	Reduced sample size	AP	AUC-ROC
1	24 915	90	0.436 ± 0.002	0.900 ± 0.000
3	3737	90	0.467 ± 0.003	0.909 ± 0.001
5	1352	90	0.477 ± 0.002	0.913 ± 0.000
15	90	90	0.512 ± 0.002	0.930 ± 0.000

The original sample size represents the sample size for each minimum visit, while the reduced sample size corresponds to the sample size for the highest minimum visit. The CI values rounded to 3 digits after the point.

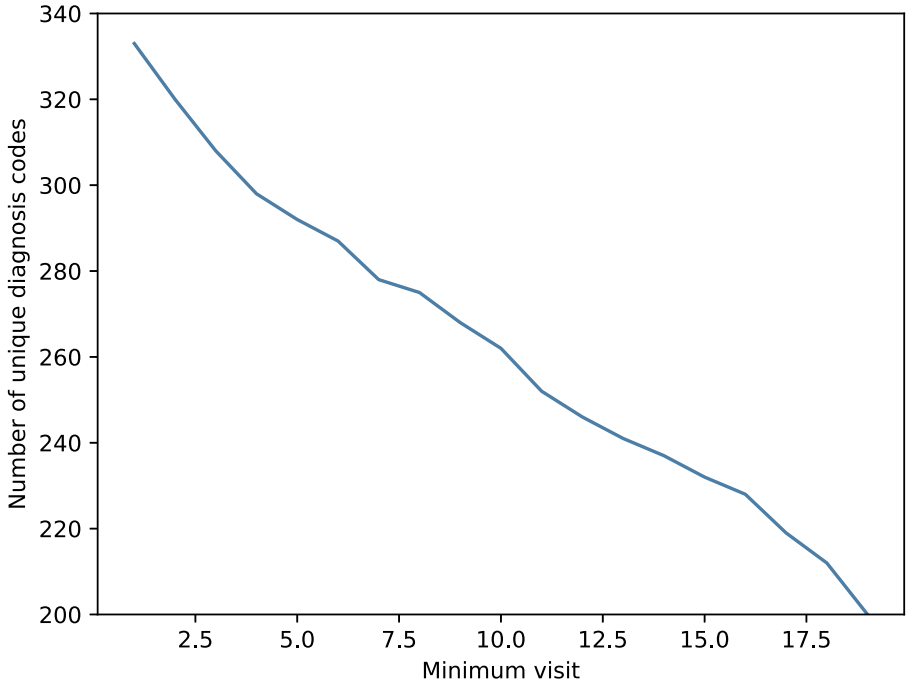

Figure 4. Number of different group of CCS diagnoses as a function of the number of the minimum visit.

We attributed this trend to a reduction in the number of available samples as the minimum visit threshold increased and there are less patients having many visits than patients with only a few visits. To support this claim, we sampled the same number of patients for each dataset. We found that the average precision and AUC-ROC of the next visit model for different thresholds (1, 3, 5, and 15) increases with the minimal number of visits while ensuring an equal number of samples across all models based on the minimum visit count of 15 (Table 1). The table substantiates our hypothesis, as it demonstrates that increasing the minimum visit number does not result in lower average precision or AUC-ROC values. Moreover, Table 1 highlights that a longer patient diagnoses history contributes to improved predictions for the next visit diagnoses of the same patient. Therefore, the decrease in performance of the next visit model when increasing the minimal number of visit threshold (when using the complete dataset) is only due to the decrease in sample size.

Diagnosis codes diversity across minimum visit thresholds

In the current work, we discussed the impact of minimum visit thresholds on the diversity of diagnoses codes and the set of concepts available for learning their embeddings. We found that as the minimum visit threshold increases, the number of different group of diagnoses decreases, indicating lower diversity (Figure 4). This observation supports our finding that for a minimum visit threshold of 1, the set of possible diagnoses and concepts available for learning their embeddings is larger compared to thresholds of 3, 5, and 15. Therefore there is a larger gap between the centralized training and the FL training for minimum visit of 1.

AUC-ROC performance

Our main evaluation metric was the average precision which is appropriate for multilabel prediction. A secondary evaluation metric we used is the AUC-ROC. Since our problem is multilabeled, we calculate the metrics separately for each label and then average the results for both AUC-ROC and average precision. We calculated the average precision and AUC-ROC using scikit-learn package in Python. Figure 5 shows the AUC-ROC performance of the different models. Similar to the average precision, our FL approach is close to the centralized training and improves the performance compared to local training, while ensuring data privacy, and also applicable for multicenter studies.

Limitations, future work, and conclusion

While our study presents promising findings, we acknowledge several limitations. Notably, we applied the FL approach outlined in this research on simulated multicenter data, rather than real-world multicenter datasets. We lacked access to a

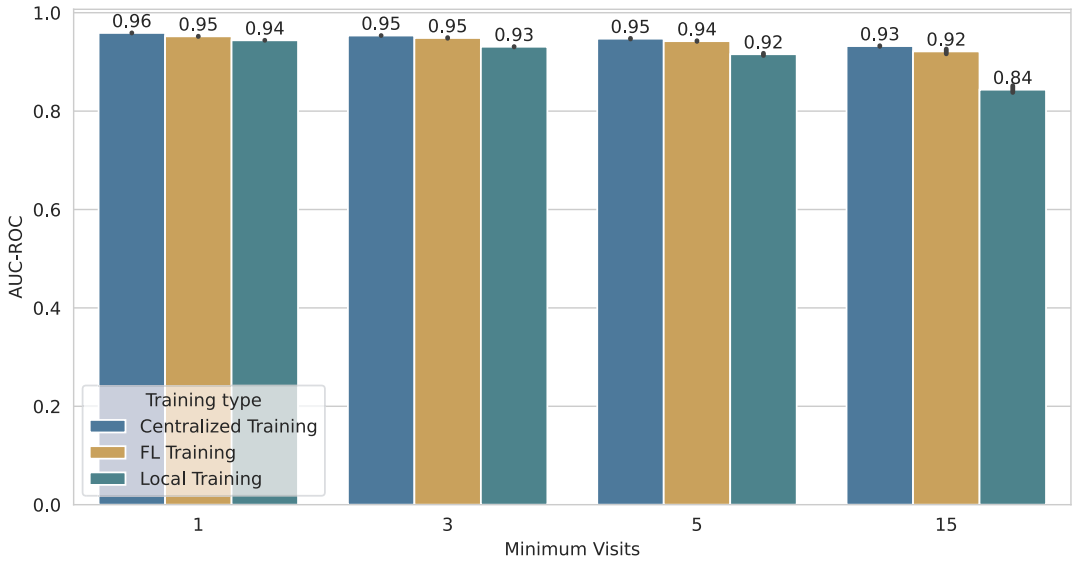

Figure 5. The AUC-ROC of each training method was evaluated for the next visit prediction task. The centralized model is referred to as non-FL training, our proposed approach is FL training, and local training involves training local models in our multicenter study. We evaluated the AUC-ROC of the models at 4 minimum visit thresholds. The AUC-ROC value appears at the top of each bar plot, and also the 95% CI. Abbreviation: FL, federated learning.

dataset with sufficient patient visit frequency for our analysis which can be used to test our approach in a federated and centralized scheme. Specifically, within the eICU-CRD dataset, the average number of visits per patient was only 1.2, making it inadequate for our purposes of predicting diagnoses in the next visit. Future work could consider combining a wider range of features, such as laboratory results and vital signs, to further improve the accuracy of the predictive models. Another interesting direction for future research would be to investigate the potential of applying alternative models, such as MedBERT,[21] to the FL approach presented in this study.

In this study, we proposed an FL approach using the FedAvg algorithm to train an MLM and a next visit prediction task, enabling the privacy of EHR data to be maintained in multicenter studies. Our FL approach achieved similar performance to the centralized model, and an improvement of 4-10 absolute percents of average precision compared to local models. This highlights the importance of our FL approach for creating a common model for multicenter studies while preserving data privacy and improving the generalizability and robustness of the model. The potential impact of our approach on clinical relevance lies in addressing variability in patient populations and enhancing the effectiveness of tailored interventions. Furthermore, our approach is general to any multicenter study and it is scalable to any number of clients, compared to local models and the centralized model baseline approaches. Our code is available at the following link: https://github.com/nadavlab/FederatedBEHRT.

Supplementary material

Supplementary material is available at *JAMIA Open* online.

Author contributions

Ofir Ben Shoham performed conceptualization, data curation, formal analysis, methodology, software validation, visualization, writing—original draft, and writing—review & editing. Nadav Rappoport performed conceptualization, investigation, methodology, resources, supervision, writing—original draft, and writing—review & editing.

Funding

None declared.

Conflicts of interest

None declared.

Data availability

In our research, we used the MIMIC-IV v2.0 dataset.[11] The data can be accessed at https://physionet.org/content/mimic-iv/2.0/.

References

1. Evans RS. Electronic health records: then, now, and in the future. *Yearb Med Inform*. 2016;Suppl 1:S48-S61.
2. Dang TK, Lan X, Weng J, Feng M. Federated learning for electronic health records. *ACM Trans Intell Syst Technol*. 2022;13:1-17.
3. Pati S, Baid U, Edwards B, et al. Federated learning enables big data for rare cancer boundary detection. *Nat Commun*. 2022;13:7346.
4. McMahan B, Moore E, Ramage D, Hampson S, y Arcas BA. Communication-efficient learning of deep networks from decentralized data. In: Artificial Intelligence and Statistics. PMLR; 2017:1273-1282.
5. Xu J, Glicksberg BS, Su C, Walker P, Bian J, Wang F. Federated learning for healthcare informatics. *J Healthc Inform Res*. 2021;5:1-19.
6. Li Y, Rao S, Solares JRA, et al. BEHRT: transformer for electronic health records. *Sci Rep*. 2020;10:7155.
7. Devlin J, Chang MW, Lee K, Toutanova K. Bert: pre-training of deep bidirectional transformers for language understanding. arXiv, arXiv:1810.04805, preprint: not peer reviewed.

8. Herrett E, Gallagher AM, Bhaskaran K, et al. Data resource profile: clinical practice research datalink (CPRD). *Int J Epidemiol*. 2015;44:827-836.
9. Choi E, Bahadori MT, Sun J, Kulas J, Schuetz A, Stewart W. Retain: an interpretable predictive model for healthcare using reverse time attention mechanism. *Ad Neural Inf Process Syst*. 2016;29.
10. Miotto R, Li L, Kidd BA, Dudley JT. Deep patient: an unsupervised representation to predict the future of patients from the electronic health records. *Sci Rep*. 2016;6:1-10.
11. Johnson AEW, Bulgarelli L, Shen L, et al. Mimic-iv, a freely accessible electronic health record dataset. *Sci Data*. 2023;10:219.
12. Si Y, Du J, Li Z, et al. Deep representation learning of patient data from electronic health records (EHR): a systematic review. *J Biomed Inform*. 2021;115:103671.
13. Douglas GM, Yu Y, Shen Y, et al. Phe2vec: automated disease phenotyping based on unsupervised embeddings from electronic health records. *J Am Med Inf Assoc*. 2020;27:1727-1735.
14. Pennington J, Socher R, Manning CD. Glove: global vectors for word representation. In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics; 2014:1532-1543.
15. Busta M, Neumann L, Matas J. Fastext: efficient unconstrained scene text detector. In: *Proceedings of the IEEE International Conference on Computer Vision*. IEEE; 2015:1206-1214.
16. Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. *Adv Neural Inf Process Syst*. 2017;30.
17. Wu CH, Xie G, Xie S, et al. Bidirectional representation learning from transformers using multimodal electronic health record data to predict depression. *IEEE J Biomed Health Inform*. 2020;24:3177-3187.
18. Liu F, Shareghi E, Meng Z, Basaldella M, Collier N. Self-alignment pretraining for biomedical entity representations. arXiv, arXiv:2010.11784, preprint: not peer reviewed.
19. Sakhovskiy A, Semenova N, Kadurin A, Tutubalina E. Biomedical entity representation with graph-augmented multi-objective transformer. In: *Findings of the Association for Computational Linguistics: NAACL 2024*. Association for Computational Linguistics; 2024:4626-4643.
20. Ronzano F, Nanavati J. Towards ontology-enhanced representation learning for large language models. arXiv, arXiv:2405.20527, preprint: not peer reviewed.
21. Rasmy L, Xiang Y, Xie Z, Tao C, Zhi D. Med-BERT: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. *NPJ Digit Med*. 2021;4:86.
22. Gao J, Wang X, Wang Y, et al. CAMP: co-attention memory networks for diagnosis prediction in healthcare. In: *2019 IEEE International Conference on Data Mining (ICDM)*. IEEE; 2019:1036-1041.
23. Ma F, Wang Y, Xiao H, et al. A general framework for diagnosis prediction via incorporating medical code descriptions. In: *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE; 2018: 1070–1075.
24. Li T, Sahu AK, Zaheer M, Sanjabi M, Talwalkar A, Smith V. Federated optimization in heterogeneous networks. In: *Proceedings of Machine Learning and Systems,* Vol. 2. 2020;429-450.
25. Boughorbel S, Jarray F, Venugopal N, Moosa S, Elhadi H, Makhlouf M. Federated uncertainty-aware learning for distributed hospital EHR data. arXiv, arXiv:1910.12191, preprint: not peer reviewed.
26. Grama M, Musat M, Muñoz-González L, Passerat-Palmbach J, Rueckert D, Alansary A. Robust aggregation for adaptive privacy preserving federated learning in healthcare. arXiv, arXiv:2009.08294, preprint: not peer reviewed.
27. Bebortta S, Tripathy SS, Basheer S, Chowdhary CL. Fedehr: a federated learning approach towards the prediction of heart diseases in IoT-based electronic health records. *Diagnostics*. 2023;13:3166.
28. Tian Y, Wan Y, Lyu L, Yao D, Jin H, Sun L. Fedbert: when federated learning meets pre-training. *ACM Trans Intell Syst Technol*. 2022;13:1-26.

JAMIA Open, 2024, 7, 1–8
https://doi.org/10.1093/jamiaopen/ooae110
Research and Applications