EMPIRICAL EVALUATIONS OF PERSONALIZED FEDERATED LEARNING ON HETEROGENEOUS ELECTRONIC HEALTH RECORDS

Yuqing Shang, Qiming Wu, Siqi Li & Di Miao
Center for Quantitative Medicine, Duke-NUS Medical School, Singapore

ABSTRACT

Beyond mobile health devices, federated learning (FL) in healthcare often occurs in cross-silo scenarios, revealing an underexplored area — the comparison of FL, including personalized FL (PFL) models with pre-existing local models. The fact that the majority of existing FL and PFL algorithms were originally designed for cross-device FL settings leaves potential room for improvements in cross-silo scenarios. Our study[1] tests several PFL frameworks on real-world heterogeneous electronic health records, and we also adapt an existing PFL framework, PerFed-Avg, to cross-silo setting by allowing personalized local epochs in clients. Results show that our modified PFL algorithm can benefit cross-silo clinical structured data, and personalizing local epochs contributes to FL model performance.

1 INTRODUCTION & METHODS

The use of federated learning (FL) in healthcare research has gained popularity, fostering collaborative research while ensuring patient privacy (Rieke et al., 2020). Although early FL adoption in medicine primarily focused on unstructured data (Crowson et al., 2022), exploration of structured data remains limited (Li et al., 2023a). Beyond differences in sample sizes and modeling strategies compared to FL on unstructured data, FL for clinical structured data is expected to yield additional meaningful results, surpassing those from pre-existing local models (Li et al., 2023a).

Personalized FL (PFL) aims to enhance algorithms through client-specific personalization (Tan et al., 2021). However, existing PFL frameworks, primarily developed for cross-device settings and validated with unstructured data (Matsuda et al., 2022), may not directly apply to many FL studies on clinical structured data. For example, PFL algorithms like FedBN (Li et al., 2021), designed for neural networks, may not align with the preferences for interpretable models, such as traditional statistical regressions (Li et al., 2023a). Therefore, empirical evaluations of PFL in cross-silo settings are crucial for understanding its applicability to clinical structured data.

To address the gap, our study conducts empirical evaluations of PFL and the baseline FL framework, FedAvg(McMahan et al., 2017), using real-world heterogeneous EHRs. Specifically, we focus on a meta-learning-based PFL algorithm, Per-FedAvg (Fallah et al., 2020) and a Morea envelopes-based PFL algorithm, pFedMe (T Dinh et al., 2020). Additionally, we propose Per-FedAvg†(Algorithm 1), a modified version of Per-FedAvg adapted for cross-silo settings by allowing different local epochs for each client. As highlighted by Wang et al. (2020) and Horváth et al. (2022), data imbalance among clients can cause inconsistency in the rate of convergence due to differences in local updates. While current PFL algorithms are unable to address this issue, we tackle it by allowing clients with smaller sample sizes to undergo additional local epochs during training. The details outlining the differences between Per-FedAvg† and Per-FedAvg are presented in Appendix A.1.

We evaluate the performance of Per-FedAvg†, the vanilla Per-FedAvg, pFedMe, and FedAvg with local models to assess whether the FL frameworks could bring benefits to participating clients. The experiments are conducted on two real-world heterogeneous EHR datasets: MIMIC-IV-ED (Johnson et al., 2023) and SGH-ED (Liu et al., 2022), with further details available in Appendix A.2. The outcome of interest in this study is inpatient mortality, and logistic regression (with a total of 15

[1] The code is publicly available at this GitHub link.

variables, details provided in Appendix A.2) is implemented as the most prevalent model in clinical structured data analysis within FL settings (Li et al., 2023a).

Algorithm 1: Per-FedAvg† Algorithm

```
Set $E = [E_1, ..., E_K]$ for K clients and initial parameter $w_0$.
for each client $k$ do
    Split client's training data into batches $\mathcal{B}$ of size $B$;
    for each local epoch $i$ from 1 to $E_k$ do
        for each batch $b$ in $\mathcal{B}$ do
            Perform Per-FedAvg SGD update;
        end
    end
    Client $k$ sends its updated parameter $w_k$ back to the server;
end
Server updates its model by averaging over received parameters: $w = \sum w_k / K$.
```

2 Results & Discussion

Table 1 summarizes top-performing models on the MIMIC dataset (client with smaller sample size), excluding ^{b}Per-FedAvg† which targets optimizing performance for SGH. Experiment details are provided in Appendix A.3. Our key observations are:

1. ^{a}Per-FedAvg†, ^{b}Per-FedAvg†, and Per-FedAvg outperform pFedMe in average testing AUROC. Notably, the optimal learning rate for pFedMe differs significantly from that of Per-FedAvg, ^{a}Per-FedAvg†, ^{b}Per-FedAvg†, and FedAvg.
2. ^{a}Per-FedAvg† outperforms Per-FedAvg on MIMIC, and ^{b}Per-FedAvg† outperforms Per-FedAvg on SGH. The improvement on MIMIC is more significant than on SGH, aligning with the observation (Appendix A.3) that SGH's AUROC values have relatively smaller fluctuations due to its larger sample size. For both ^{a}Per-FedAvg† and ^{b}Per-FedAvg†, the performance increase on the target client is accompanied by a drop on the other client, indicating a trade-off between personalization and generalization.
3. Both ^{b}Per-FedAvg† and Per-FedAvg benefit the MIMIC dataset (the target client), as demonstrated by comparisons with local models. In contrast, FedAvg and pFedMe do not exhibit such benefits.

Table 1: Performance of PFL and baseline models along with their fine-tuned hyperparameters. lr: learning rate; T: global communication rounds; ($E1$,$E2$): (MIMIC local epoch, SGH local epoch).

	Test AUROC(%)			**Parameters**		
	MIMIC test	SGH test	Average	lr	T	($E1$,$E2$)
Local MIMIC	73.45%	81.83%	77.64%	-	-	-
Local SGH	71.57%	86.15%	78.86%	-	-	-
FedAvg	72.83%	86.19%	79.51%	0.5	30	(10,10)
pFedMe	71.69%	85.86%	78.77%	0.005	30	(10,10)
Per-FedAvg	73.51%	86.00%	79.76%	1	11	(10,10)
^{a}**Per-FedAvg†**	**74.14**%	85.74%	**79.94%**	1	10	(13,10)
^{b}**Per-FedAvg†**	73.14%	**86.20**%	79.67%	1	10	(10,4)

In this study, we investigated PFL's effectiveness on clinical structured data in cross-silo FL settings. Our findings show that both Per-FedAvg and our proposed method outperform pFedMe. Notably, our Per-FedAvg† exhibits enhanced performance on both clients, especially on the client with a smaller sample size.

Our results highlight the importance of increasing personalization through leveraging more local information to consistently bring benefits to specific participating clients in cross-silo PFL. We anticipate that our empirical findings could offer insights and serve as a source of inspiration for future research directions in cross-silo PFL.

URM Statement

The authors acknowledge that all authors of this work meet the URM criteria of the ICLR 2024 Tiny Papers Track.

References

Matthew G. Crowson, Dana Moukheiber, Aldo Robles Arévalo, Barbara D. Lam, Sreekar Mantena, Aakanksha Rana, Deborah Goss, David W. Bates, and Leo Anthony Celi. A systematic review of federated learning applications for biomedical data. *PLOS Digital Health*, 1(5):e0000033, May 2022. ISSN 2767-3170. doi: 10.1371/journal.pdig.0000033. URL `http://dx.doi.org/10.1371/journal.pdig.0000033`.

Alireza Fallah, Aryan Mokhtari, and Asuman Ozdaglar. Personalized federated learning with theoretical guarantees: A model-agnostic meta-learning approach. *Advances in Neural Information Processing Systems*, 33:3557–3568, 2020.

Samuel Horváth, Maziar Sanjabi, Lin Xiao, Peter Richtárik, and Michael Rabbat. Fedshuffle: Recipes for better use of local work in federated learning. *arXiv preprint arXiv:2204.13169*, 2022.

Alistair EW Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, et al. Mimic-iv, a freely accessible electronic health record dataset. *Scientific data*, 10(1):1, 2023.

Siqi Li, Pinyan Liu, Gustavo G Nascimento, Xinru Wang, Fabio Renato Manzolli Leite, Bibhas Chakraborty, Chuan Hong, Yilin Ning, Feng Xie, Zhen Ling Teo, Daniel Shu Wei Ting, Hamed Haddadi, Marcus Eng Hock Ong, Marco Aurélio Peres, and Nan Liu. Federated and distributed learning applications for electronic health records and structured medical data: a scoping review. *Journal of the American Medical Informatics Association*, 30(12):2041–2049, 08 2023a. ISSN 1527-974X.

Siqi Li, Yilin Ning, Marcus Eng Hock Ong, Bibhas Chakraborty, Chuan Hong, Feng Xie, Han Yuan, Mingxuan Liu, Daniel M. Buckland, Yong Chen, and Nan Liu. Fedscore: A privacy-preserving framework for federated scoring system development. *Journal of Biomedical Informatics*, 146: 104485, 2023b. ISSN 1532-0464. doi: https://doi.org/10.1016/j.jbi.2023.104485. URL `https://www.sciencedirect.com/science/article/pii/S153204642300206X`.

Xiang Li, Kaixuan Huang, Wenhao Yang, Shusen Wang, and Zhihua Zhang. On the convergence of fedavg on non-iid data. *arXiv preprint arXiv:1907.02189*, 2019.

Xiaoxiao Li, Meirui Jiang, Xiaofei Zhang, Michael Kamp, and Qi Dou. Fedbn: Federated learning on non-iid features via local batch normalization. *arXiv preprint arXiv:2102.07623*, 2021.

Nan Liu, Feng Xie, Fahad Javaid Siddiqui, Andrew Fu Wah Ho, Bibhas Chakraborty, Gayathri Devi Nadarajan, Kenneth Boon Kiat Tan, and Marcus Eng Hock Ong. Leveraging large-scale electronic health records and interpretable machine learning for clinical decision making at the emergency department: Protocol for system development and validation. *JMIR Research Protocols*, 11(3): e34201, March 2022. ISSN 1929-0748. doi: 10.2196/34201. URL `http://dx.doi.org/10.2196/34201`.

Koji Matsuda, Yuya Sasaki, Chuan Xiao, and Makoto Onizuka. An empirical study of personalized federated learning. *arXiv preprint arXiv:2206.13190*, 2022.

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

Nicola Rieke, Jonny Hancox, Wenqi Li, Fausto Milletarì, Holger R. Roth, Shadi Albarqouni, Spyridon Bakas, Mathieu N. Galtier, Bennett A. Landman, Klaus Maier-Hein, Sébastien Ourselin, Micah Sheller, Ronald M. Summers, Andrew Trask, Daguang Xu, Maximilian Baust, and M. Jorge Cardoso. The future of digital health with federated learning. *npj Digital Medicine*, 3(1), September 2020. ISSN 2398-6352. doi: 10.1038/s41746-020-00323-1. URL http://dx.doi.org/10.1038/s41746-020-00323-1.

Canh T Dinh, Nguyen Tran, and Josh Nguyen. Personalized federated learning with moreau envelopes. *Advances in Neural Information Processing Systems*, 33:21394–21405, 2020.

Alysa Ziying Tan, Han Yu, Lizhen Cui, and Qiang Yang. Towards personalized federated learning. *CoRR*, abs/2103.00710, 2021. URL https://arxiv.org/abs/2103.00710.

Jianyu Wang, Qinghua Liu, Hao Liang, Gauri Joshi, and H Vincent Poor. Tackling the objective inconsistency problem in heterogeneous federated optimization. *Advances in neural information processing systems*, 33:7611–7623, 2020.

Feng Xie, Jun Zhou, Jin Wee Lee, Mingrui Tan, Siqi Li, Logasan S/O Rajnthern, Marcel Lucas Chee, Bibhas Chakraborty, An-Kwok Ian Wong, Alon Dagan, Marcus Eng Hock Ong, Fei Gao, and Nan Liu. Benchmarking emergency department prediction models with machine learning and public electronic health records. *Scientific Data*, 9(1), October 2022a. ISSN 2052-4463. doi: 10.1038/s41597-022-01782-9. URL http://dx.doi.org/10.1038/s41597-022-01782-9.

Feng Xie, Jun Zhou, Jin Wee Lee, Mingrui Tan, Siqi Li, Logasan S/O Rajnthern, Marcel Lucas Chee, Bibhas Chakraborty, An-Kwok Ian Wong, Alon Dagan, et al. Benchmarking emergency department prediction models with machine learning and public electronic health records. *Scientific Data*, 9(1):658, 2022b.

A Appendix

A.1 Algorithm of Per-FedAvg†

The primary differentiation between Per-FedAvg† and Per-FedAvg arises from two key aspects: firstly, the adaptation of local epochs, where the local epoch used by each client is allowed to be different and fine-tuned to maximize the leverage of local information; and secondly, the omission of the client random selection process in vanilla Per-FedAvg, as each client in cross-silo clinical settings is anticipated to participate and benefit from FL. This simplification not only contributes to more stable training process but also yields distinct theoretical outcomes compared to scenarios involving partial client participation, as explored in previous work (Li et al., 2019).

A.2 Details of Datasets

MIMIC-IV-ED is an open-source dataset of emergency department (ED) admissions at Beth Israel Deaconess Medical Center between 2011 and 2019. We first construct a master dataset following the pipeline proposed by Xie et al. (2022a). The dataset is then filtered to include only ED admissions of Asian patients who were at least 21 years old. Observations with missing values are removed, resulting in a final cohort of 9071 admissions. The SGH-ED dataset is a private dataset collected in the ED of Singapore General Hospital and extracted from the SingHealth Electronic Health Intelligence System. A waiver of consent was granted for EHRs data collection and retrospective analysis, and the study has been approved by the Singapore Health Services' Centralized Institutional Review Board, with all data de-identified. The dataset is filtered to include only ED admissions of adult Chinese patients in 2019. Observations with missing values are also removed, resulting in a final cohort of 81,110 admissions.

The outcome of interest in this study is inpatient mortality. Based on existing works (Li et al., 2023b; Xie et al., 2022b), we include 15 candidate variables commonly considered to be associated with inpatient mortality as predictors. Specifically, we have included two demographic variables: age and gender; five vital signs: pulse (beats/min), respiration (times/min), peripheral capillary oxygen saturation ($\mathrm{SpO_2}$; %), diastolic blood pressure (mm Hg), systolic blood pressure (mm Hg); and eight comorbidities: myocardial infarction, congestive heart failure, peripheral vascular disease, stroke, chronic pulmonary disease, rheumatic disease, paralysis and kidney disease. Both datasets are randomly split into training and testing sets at a ratio of 6:4.

A.3 Experiment Design

In this study, the batch size B is consistently set to 128. The learning rate is chosen from the range [0.005, 0.01, 0.05, 0.1, 0.5, 1] for both FL and PFL methods, with each method undergoing 30 global rounds for every selected learning rate. The local epochs are fixed at 10 for both clients in FedAvg, pFedMe, and vanilla Per-FedAvg.

^{a}Per-FedAvg† and ^{b}Per-FedAvg† differ in their optimization goals: ^{a}Per-FedAvg† aims to optimize performance on client MIMIC, while ^{b}Per-FedAvg† aims to optimize for client SGH. To find the best ($E1$,$E2$) for Per-FedAvg†, we fix other hyperparameters (lr=1, T=10, B=128) and conduct the following fine-tuning process. First, we fix $E2$=10 and experiment with ($E1$,$E2$) pairs: (5, 10), (7, 10), (9, 10), (10, 10), (11, 10), (12, 10), (13, 10), (15, 10), (20, 10), (25, 10), (30, 10), (40, 10), (50, 10), (70, 10), (100, 10), (140, 10). Second, we fix $E1$=10 and experiment with ($E1$, $E2$) pairs: (10, 2), (10, 3), (10, 4), (10, 5), (10, 6), (10, 7), (10, 8), (10, 10), (10, 12), (10, 15), (10, 20), (10, 25), (10, 35), (10, 44), (10, 50), (10, 70).

Figure A.1 illustrates that the parameter combination (13, 10) yields the best performance on MIMIC test data, as reported for ^{a}Per-FedAvg† in Table 1. Similarly, as depicted in Figure A.2, the parameter combination (10, 4) delivers the optimal performance on SGH test data, as reported for ^{b}Per-FedAvg† in Table 1.

Figure A.1: Performance of Per-FedAvg† across $(E1, E2)$ with a fixed value of $E2 = 10$.

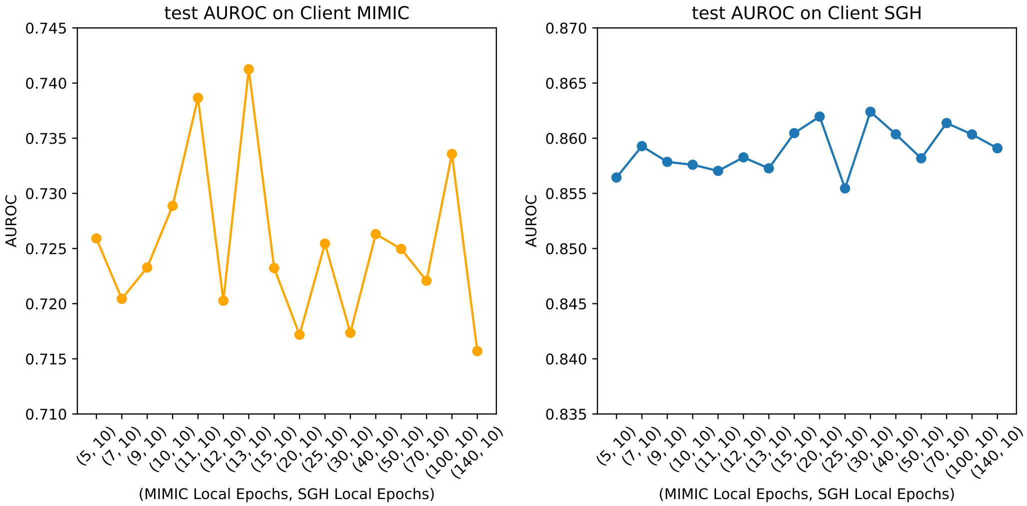
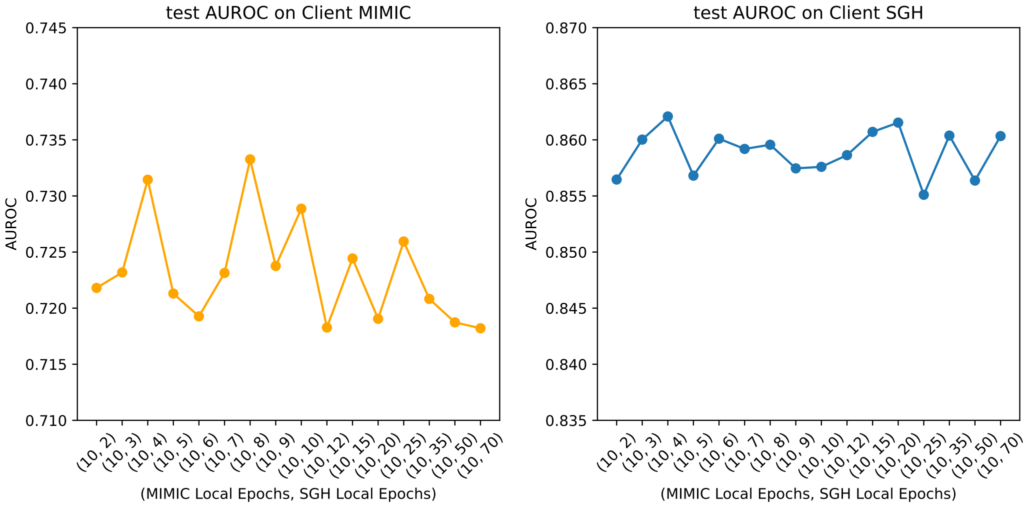


Figure A.2: Performance of Per-FedAvg† across $(E1, E2)$ with a fixed value of $E1 = 10$.


test AUROC on Client MIMIC
AUROC
(10, 2)
(10, 3)
(10, 4)
(10, 5)
(10, 6)
(10, 7)
(10, 8)
(10, 9)
(10, 10)
(10, 12)
(10, 15)
(10, 20)
(10, 25)
(10, 35)
(10, 50)
(10, 70)
(MIMIC Local Epochs, SGH Local Epochs)
test AUROC on Client SGH
AUROC
(10, 2)
(10, 3)
(10, 4)
(10, 5)
(10, 6)
(10, 7)
(10, 8)
(10, 9)
(10, 10)
(10, 12)
(10, 15)
(10, 20)
(10, 25)
(10, 35)
(10, 50)
(10, 70)
(MIMIC Local Epochs, SGH Local Epochs)