Best Practice

DIGITAL HEALTH

Digital Health
Volume 8: 1–5

DOI: 10.1177/20552076221134455
journals.sagepub.com/home/dhj
SAGE

Federated learning for preserving data privacy in collaborative healthcare research

Tyler J Loftus[1,2], Matthew M Ruppert[2,3], Benjamin Shickel[2,4], Tezcan Ozrazgat-Baslanti[2,3], Jeremy A Balch[1,2], Philip A Efron[1], Gilbert R Upchurch Jr.[1], Parisa Rashidi[2,5], Christopher Tignanelli[6], Jiang Bian[7] and Azra Bihorac[1,2]

Abstract

Generalizability, external validity, and reproducibility are high priorities for artificial intelligence applications in healthcare. Traditional approaches to addressing these elements involve sharing patient data between institutions or practice settings, which can compromise data privacy (individuals' right to prevent the sharing and disclosure of information about themselves) and data security (simultaneously preserving confidentiality, accuracy, fidelity, and availability of data). This article describes insights from real-world implementation of federated learning techniques that offer opportunities to maintain both data privacy and availability via collaborative machine learning that shares knowledge, not data. Local models are trained separately on local data. As they train, they send local model updates (e.g. coefficients or gradients) for consolidation into a global model. In some use cases, global models outperform local models on new, previously unseen local datasets, suggesting that collaborative learning from a greater number of examples, including a greater number of rare cases, may improve predictive performance. Even when sharing model updates rather than data, privacy leakage can occur when adversaries perform property or membership inference attacks which can be used to ascertain information about the training set. Emerging techniques mitigate risk from adversarial attacks, allowing investigators to maintain both data privacy and availability in collaborative healthcare research. When data heterogeneity between participating centers is high, personalized algorithms may offer greater generalizability by improving performance on data from centers with proportionately smaller training sample sizes. Properly applied, federated learning has the potential to optimize the reproducibility and performance of collaborative learning while preserving data security and privacy.

Keywords

Federated learning, deep learning, data, security, privacy

Submission date: 26 September 2022; Acceptance date: 5 October 2022

Introduction

Generalizability and external validity are important, and often missing, elements of artificial intelligence applications in healthcare. Without generalizability and external validity, it is difficult to establish reproducibility, which is an essential element of any scientific investigation that is often missing.[1] Reproducibility has even greater importance for artificial intelligence applications in healthcare because lack of trust is a major barrier to their clinical implementation. Patients, clinicians, and investigators may be willing to use an artificial intelligence algorithm that is well-validated, even if its inner

[1]Department of Surgery, University of Florida, Gainesville, FL, USA
[2]University of Florida, Intelligent Critical Care Center, Gainesville, FL, USA
[3]Department of Medicine, University of Florida, Gainesville, FL, USA
[4]Department of Biomedical Engineering, University of Florida, Gainesville, FL, USA
[5]Departments of Biomedical Engineering, Computer and Information Science and Engineering, and Electrical and Computer Engineering, University of Florida, Gainesville, FL, USA
[6]Department Surgery, University of Minnesota, Minneapolis, MN, USA
[7]Department of Health Outcomes and Biomedical Informatics, University of Florida, Gainesville, FL, USA

Corresponding author:
Azra Bihorac, University of Florida, Intelligent Critical Care Center, Gainesville, FL, USA.
Email: abihorac@ufl.edu; Twitter: @AzraBihorac

workings are somewhat mysterious, just as they trust medications that have proven efficacy and unclear mechanisms of action. In addition, a reproducible artificial intelligence application that is shared with academic communities can be tuned and optimized over time, improving its performance.

For artificial intelligence applications in healthcare, traditional approaches to establishing generalizability and external validity involve sharing patient data between institutions or practice settings, which can compromise data privacy (individuals' right to prevent the sharing and disclosure of information about themselves) and data security (simultaneously preserving confidentiality, accuracy, fidelity, and availability of data). Sharing data between institutions is further complicated by different hospitals using different electronic health records.[2]

Federated learning has the potential to improve reproducibility by leveraging data from multiple settings to produce generalizable results while maintaining data privacy and security via collaborative machine learning knowledge sharing without sharing the underlying data (Figure 1). Acquiring more, diverse data for artificial intelligence training datasets is a positive-sum game, as evident in artificial intelligence models in general and in federated learning models that perform better on local datasets than similar models using local data exclusively.[3]

Despite the potential advantages of federated learning for improving the reproducibility, generalizability, and performance of artificial intelligence applications in healthcare without compromising security and privacy, federated learning is rarely reported in published literature, which may be partially attributable to a lack of understanding of the logistical challenges and potential solutions in real-world implementation. This article describes the foundational principles of federated learning and offers insights from real-world implementation of federated learning model to diagnose COVID-19 using chest radiographs from five international healthcare systems, four in the United States and one in Spain, including 87,956 patients, 24,100 of whom were COVID positive.

Dataset preparation

Individual healthcare centers, even those within the same healthcare network using the same electronic medical record system, have varying schemas of storing data that were created to support the various clinical workflows specific to each site. Therefore, mapping data to a common data model (CDM) is a prerequisite for any inter-institutional project. CDMs define guidelines for consistent organization of data through standardized vocabularies, naming, and indexing methods. To maximize reproducibility and interoperability, we suggest the mapping be done to a well-established and widely adopted CDM such as The National Patient-Centered Clinical Research Network (PCORnet) or the Observational Medical Outcomes Partnership (OMOP) instead of a proprietary, project-specific format. Under these circumstances, federated learning datasets can contain nearly any multi-modal data type that can be parsed by deep learning models. For projects with few variables and little ambiguity regarding variable semantics, a project specific CDM shared among participating institutions may allow for more rapid development, which is why it was selected for COVID-19 chest radiograph project that used nine unambiguous variables, but this comes at the cost of slowing its adoption to other centers due to the need to create project specific datasets.

In order to develop and/or execute a model, various preprocessing steps are needed to organize the necessary data elements (e.g. variable selection, resampling), convert them to the required types (e.g. concept extraction, one-hot encoding, tokenization), and perform any required standardization steps (e.g. outlier removal, normalization, imputation). Variable selection is a crucial step in any project but is paramount in any multicenter project to ensure that each variable is available, consistently populated, and easily accessible across all institutions. For the COVID-19 chest radiograph project, we used age, sex, race, ethnicity, COVID-19 status (positive/negative), intensive care unit admission (yes/no), requirement for mechanical ventilation (yes/no), hospital mortality (yes/no), and admission chest radiograph images. In addition to simplicity, these variables have the added advantages that they contain none of the 18 HIPPA identifiers that permitted the sharing of the data for performance comparisons between centralized and distributed learning methods, without a data use agreement, and facilitated expedited IRB approval (within 8 days of submission). By starting with a CDM, these preprocessing steps were readily shared between sites to further ensure reproducibility.

Model development

The selection of model type in federated learning is primarily dependent on the prediction/classification task, however; gradient based learning models, such as neural networks are commonly used due to the ability to easily aggregate gradients between individual centers. In the COVID-19 chest radiograph project, we used a DenseNet (convolutional neural network) model pretrained on ImageNet with rectified linear activation functions, a range of layers in each pooling block (6, 12, 24, and 16), and a growth rate of 32 (adding 32 filters for each layer), which was later tuned for COVID-19 detection using federated learning.

Once a model type is determined, then the federated learning algorithm can be selected. This is the algorithm responsible for coordinating the training process through aggregating knowledge from each site (often in the form of coefficients/gradients) and distributing model updates. Independent and identically distributed (iid) datasets can utilize a popular algorithm FedAvg that is fast and efficient,

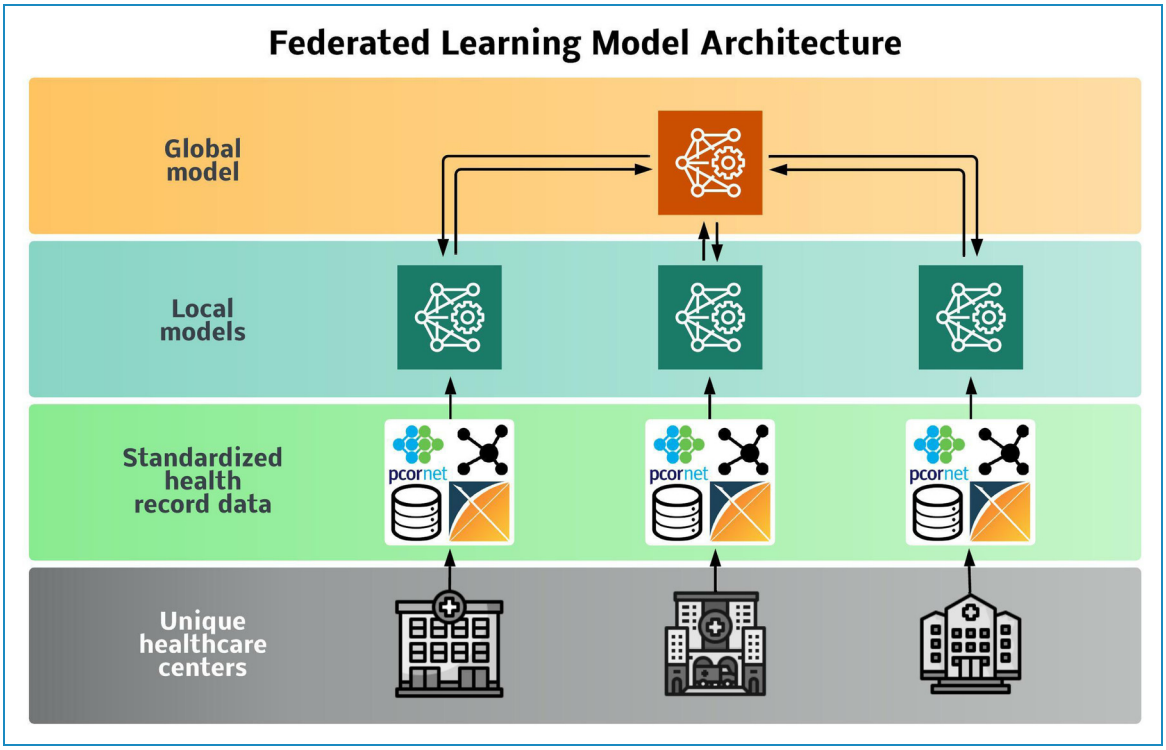

Figure 1. Federated learning model architecture. Individual healthcare centers standardize their health record data using common data models. Local models then train at each center. Local model parameters and coefficients are shared with a global model, which consolidates the local model parameters and coefficients, and then shares them with local models.

but often performs poorly when center-to-center data is heterogenous.[4–6] To address this issue, several other personalized federated learning algorithms (e.g. FedAMP, FedBN, FedProx) were created to address the data heterogeneity issue, by tailoring updates by weighting contributions from more similar models/datasets to the target model more than dissimilar ones.

In the COVID-19 chest radiograph project, in the first experiments, local model training and testing was performed at one site—the largest site, which included 45,513 patients, 3997 of whom were COVID positive. The local model had an internal AUROC of 0.94 and AUPRC of 0.85, but during testing on two separate external validation datasets, AUROC fell to 0.56 and 0.67, and AUPRC fell to 0.57 and 0.60. Next, the single-institution model was compared with federated learning models. We found the FedAvg to perform well when testing on data from centers that contributed the greatest proportions of the training data, and FedBN (in which all weights are shared and aggregated except from batch normalization layers) outperformed FedAMP (in which each local model has a dedicated global model) and FedProx (which adds a proximal term to the objective of the local update) in addressing data heterogeneity on real-world imaging data. On two external validation datasets, FedAmp had AUROC 0.71 and 0.62, FedBN had AUROC 0.78 and 0.70, and FedProx had AUROC 0.78 and 0.67. In particular, FedBN maintained reasonably strong predictive performance when testing on data from centers with proportionately smaller training sample sizes, while other federated learning algorithms, especially FedAvg, performed poorly when testing on data from centers with small training samples. We performed experiments using each of these algorithms to better understand their effects on predictive performance.

Predictive performance improvement with federated learning

Artificial intelligence algorithms are data driven, learning from examples; thus, in many cases, providing more examples facilitates greater learning. One of the primary mechanisms by which having more examples improves learning is greater representation of rare cases. If a rare constellation of variables has a rare association with the outcome of interest and there are few examples of this phenomenon in the training data, then it is unlikely that the algorithm will learn to detect those constellations of variables and associate them with the outcome in the validation and test sets. With more total examples, the algorithm has additional opportunities to learn those rare associations and make accurate predictions when confronted with similar circumstances in the future.

Another one of the major explanations to the improvement of models in the context of federated learning is related to the growing complexity of AI models and under training. With greatly improved access to high performance computing platforms, it is now feasible to create models with millions or even billions of parameters that require a tremendous number of training samples in order to achieve their maximum potential.

One might reasonably hypothesize that a local model that trains and tests exclusively on local data would learn associations between inputs and outcomes with greater accuracy than a model that trains and tests on heterogeneous data from other centers that have different patients, different healthcare providers, and different healthcare delivery modes and protocols, but the opposite is often true. In some cases, federated learning models perform better on local datasets than similar models using local data exclusively.[3] Presumably, this is because federated models learn more accurate representations from a greater number of instances, especially rare instances that are underrepresented in small datasets. In an investigation using chest radiographs to diagnose COVID-19 on a federated learning platform, we found that the area under the receiver operating characteristic curve improved by 3% and area under the precision-recall curve improved by 8% compared with local model training. Such observations are likely attributable to the enhanced ability of federated learning models to learn from more examples, even if those examples are heterogeneous and different than those in local datasets. To solve the heterogeneity problem, we recommend using a CDM at each institution to map each contributing dataset to a well-established and widely adopted CDM such as The National Patient-Centered Clinical Research Network (PCORnet) or the Observational Medical Outcomes Partnership (OMOP).

Privacy leakage in federated learning

Although the sharing of knowledge in the form of model gradients or coefficients is far more secure than the sharing of the underlying data, there is a small, but limited risk posed by malicious actors performing what are known as privacy attacks. Privacy attacks can be categorized on a spectrum from "white box" (an adversary has full access to the model and/or dataset from which the training data was drawn) to "black box" (no knowledge of model (architecture nor parameters) or training data). Although most artificial intelligence models are vulnerable to such attacks, federated learning increases the opportunity for such attacks due to multiple centers having "white box" access to the model and the communication of gradients between local institutions and the coordinating center which could be intercepted.

Patient-level privacy leaks occur through membership inference attacks (i.e. adversaries infer whether a given patient or set of patients belong to the training data)[7,8] and reconstruction attacks (i.e. adversaries reconstruct input from model output).[9–11] Privacy attacks can be mitigated through introduction of stochastic noise (e.g. differential privacy), rounding, k-anonymity, and/or l-diversity, however; all of these strategies have some impact on utility.

Intentionally rounding or embedding noise into input data, the model itself, or model outputs are common methods of increasing privacy.[12] Although models can internally account for the introduction of the noise, their performance will decrease in proportion to the amount of additional privacy protection provided by the modification. The binning of input data to achieve k-anonymity and/or l-diversity is another common method of protecting privacy such as grouping age into deciles (0–9 years, 10–19 years, 20–29 years) which can also be helpful in preserving privacy. Depending on the bin width and the relationship between the binned variables, there may be a lesser tradeoff between performance and privacy for this method compared to rounding or the addition of noise. Lastly, the encryption of updates between the local and central model is a method of reducing the vulnerability of such attacks without compromising performance.

It is important to remember that in order for these types of attacks to be successful they require a large amount of external information (e.g. information on specific patients to test against, access to the data source where the training data was procured from, knowledge of model architecture/parameters) and the ability to query the model thousands to millions of times. Lastly, the information that an adversary can possibly extract from such an attack is limited to whatever information was fed into the model, which is highly unlikely to contain hard identifiers such as birthdate, social security number, dates of service, medical record number, full name, etc., thus, even with a lot of external information about individuals, it can be nearly impossible to re-identify a particular patient or group of patients. To further minimize the threat of adversarial attacks, data privacy can be protected at the point of data collection via clustering-based anonymization.[13]

Hardware and software considerations for computational and space complexity

One of the advantages of federated learning is distributing the training burden in terms of computing resources amongst all the participating sites. Recreating the same computing environment in different locations and/or operating systems can be difficult even with package managers such as anaconda, thus we recommend the use of virtualized container environments, such as docker, that allow identical software stacks to run on any computing platform that supports docker technology. Furthermore, if cloud computing resources are used, such as Google Cloud Platform (GCP), Amazon Web Services (AWS), or Microsoft Azure, we recommended selecting one cloud

provider to simplify the process of getting each site up and running and communicating with the coordinating center. For the COVID-19 chest radiograph project we used AWS instances equipped with single or multiple V100 GPUs (the p3 instance family), containerized using Docker, with the NGC Clara Train v4.0 as the base image.

Conclusions

Federated learning can produce generalizable and reproducible artificial intelligence algorithms in healthcare. CDMs and shared preprocessing steps can improve the efficiency and interoperability of federated learning platforms. Although federated learning models are not immune to privacy leakage, emerging techniques mitigate risk for adversarial attacks. By incorporating more examples from more diverse training datasets, federated learning can outperform local models on local test data. The commonly used FedAvg algorithm may perform poorly when data heterogeneity between participating centers is high. Personalized algorithms can adapt to data heterogeneity by finding models that adapt to each dataset rather than training a shared, global model directly, and maintain stable performance on test data from centers with proportionately smaller training sample sizes. When these principles are properly applied, federated learning has the potential to optimize the reproducibility and performance of collaborative learning while preserving data security and privacy.

Acknowledgements: The authors acknowledge the University of Florida Integrated Data Repository for its role in building datasets for federated learning projects.

Conflict of interest: The authors have no conflicts of interest to declare.

Contributions: TJL, MMR, and AB designed the study. TJL and MMR drafted the manuscript. BS, TOB, JAB, PAE, GRU, PR, and AB provided critical revisions. All authors contributed to interpretation of results and approved the final version of the manuscript.

Ethical approval: Institutional Review Board approval was obtained at each of the participating institutions.

Funding: The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: TJL was supported by the National Institute of General Medical Sciences (NIGMS) of the National Institutes of Health under Award Number K23GM140268 and by the Thomas H. Maren Fund. T.O.B. was supported by the National Institute of Diabetes and Digestive and Kidney Diseases of the National Institutes of Health grant K01DK120784, R01GM110240 from the National Institute of General Medical Sciences, and by UF Research AWD09459 and the Gatorade Trust, University of Florida. PR was supported by National Science Foundation CAREER award 1750192, P30AG028740 and R01AG05533 from the NIA, 1R21EB027344 from the National Institute of Biomedical Imaging and Bioengineering (NIBIB), and R01GM110240 from the NIGMS. AB was supported by R01GM110240 from the NIGMS and 1R21EB027344 from the NIBIB. This work was supported in part by the National Center for Advancing Translational Sciences and Clinical and Translational Sciences Award to the University of Florida UL1TR000064. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health

Guarantor: The corresponding author, AB, is the guarantor for this article.

ORCID iD: Tyler J Loftus https://orcid.org/0000-0001-5354-443X

References

1. Baker M. 1,500 Scientists lift the lid on reproducibility. *Nature* 2016; 533: 452–454.
2. Xu J, Glicksberg BS, Su C, et al. Federated learning for healthcare informatics. *J Healthc Inform Res* 2021; 5: 1–19.
3. Dayan I, Roth HR, Zhong A, et al. Federated learning for predicting clinical outcomes in patients with COVID-19. *Nat Med* 2021; 27: 1735–1743.
4. Sun T, Li D and Wang B. Decentralized federated averaging. arXiv:2104.11375.
5. Li T, Sahu AK, Zaheer M, et al. Federated optimization in heterogeneous networks. arXiv:1812.06127.
6. Zhao Y, Li M, Lai L, et al. Federated learning with non-iid data. arXiv:1806.00582.
7. Melis L, Song C, Cristofaro ED, et al. Exploiting unintended feature leakage in collaborative learning. In: 2019 IEEE symposium on security and privacy (SP) 19-23 May 2019, pp.691–706.
8. Nasr M, Shokri R and Houmansadr A. Comprehensive privacy analysis of deep learning: passive and active white-box inference attacks against centralized and federated learning. 2019, p. 739–753.
9. Wei W, Liu L, Loper M, et al. A Framework for Evaluating Gradient Leakage Attacks in Federated Learning. *arXiv preprint arXiv:200410397*.
10. Hitaj B, Ateniese G and Perez-Cruz F. Deep models under the GAN: information leakage from collaborative deep learning. Proceedings of the 2017 ACM SIGSAC conference on computer and communications security. Dallas, Texas, USA: Association for Computing Machinery, 2017, P. 603–618.
11. Wang Z, Song M, Zhang Z, et al. Beyond inferring class representatives: user-level privacy leakage from federated learning. In: IEEE Conference on computer communications INFOCOM 29 April-2 May 2019, pp.2512–2520.
12. Li JC, Meng Y, Ma LC, et al. A federated learning based privacy-preserving smart healthcare system. *Ieee T Ind Inform* 2022; 18: 2021–2031.
13. Onesimu JA, Karthikeyan J and Sei Y. An efficient clustering-based anonymization scheme for privacy-preserving data collection in IoT based healthcare services. *Peer Peer Netw Appl* 2021; 14: 1629–1649.