Methods and Impact for Using Federated Learning to Collaborate on Clinical Research

Alexander T. M. Cheung, MBHL [*]
Mustafa Nasir-Moin, AB [*]
Young Joon (Fred) Kwon, PhD[*]
Jiahui Guan, PhD[‡]
Chris Liu, BS[*]
Lavender Jiang, BS[*§]
Christian Raimondo, BS[*]
Silky Chotai, MD[‖]
Lola Chambless, MD[‖]
Hasan S. Ahmad, BS[¶]
Daksh Chauhan, BS[¶]
Jang W. Yoon, MD, MSc[¶]
Todd Hollon, MD[#]
Vivek Buch, MD[**]
Douglas Kondziolka, MD[*]
Dinah Chen, MD[††]
Lama A. Al-Aswad, MD, MPH[††]
Yindalon Aphinyanaphongs, MD, PhD[‡‡]
Eric Karl Oermann, MD[*§§§]

(Continued on next page)

Correspondence:
Eric Karl Oermann, MD,
Department of Neurosurgery,
NYU Langone Health,
530 First Ave, Skirball, 8R,
New York, NY 10016, USA.
Email: eric.oermann@nyulangone.org

Received, June 5, 2022.
Accepted, August 20, 2022.
Published Online, November 8, 2022.

BACKGROUND: The development of accurate machine learning algorithms requires sufficient quantities of diverse data. This poses a challenge in health care because of the sensitive and siloed nature of biomedical information. Decentralized algorithms through federated learning (FL) avoid data aggregation by instead distributing algorithms to the data before centrally updating one global model.

OBJECTIVE: To establish a multicenter collaboration and assess the feasibility of using FL to train machine learning models for intracranial hemorrhage (ICH) detection without sharing data between sites.

METHODS: Five neurosurgery departments across the United States collaborated to establish a federated network and train a convolutional neural network to detect ICH on computed tomography scans. The global FL model was benchmarked against a standard, centrally trained model using a held-out data set and was compared against locally trained models using site data.

RESULTS: A federated network of practicing neurosurgeon scientists was successfully initiated to train a model for predicting ICH. The FL model achieved an area under the ROC curve of 0.9487 (95% CI 0.9471-0.9503) when predicting all subtypes of ICH compared with a benchmark (non-FL) area under the ROC curve of 0.9753 (95% CI 0.9742-0.9764), although performance varied by subtype. The FL model consistently achieved top three performance when validated on any site's data, suggesting improved generalizability. A qualitative survey described the experience of participants in the federated network.

CONCLUSION: This study demonstrates the feasibility of implementing a federated network for multi-institutional collaboration among clinicians and using FL to conduct machine learning research, thereby opening a new paradigm for neurosurgical collaboration.

KEY WORDS: Artificial intelligence, Federated learning, Intracranial hemorrhage, Machine learning

Neurosurgery 00:1–8, 2022 https://doi.org/10.1227/neu.0000000000002198

One of the fundamental requirements for biomedical and health care research, particularly when it involves artificial intelligence technologies, is obtaining sufficient quantities of diverse data to obtain generalizable, equitable, and effective results.[1,2] Historically, this has led to the creation of large multicenter consortiums[3,4] or restricted research to sufficiently large academic medical centers.[5-8] This paradigm for collaborative research is built around the concept of a centralized data repository that stores samples to facilitate subsequent analyses of the aggregated data. However, a new technological approach to decentralized algorithms—federated learning (FL)[9]—has challenged this paradigm. In FL, a federated network facilitates the training of algorithms on distributed data by bringing the learning algorithm to participating sites, thereby eliminating the need to transfer files to a central repository for storage and analysis.[10] In theory, by freeing research groups of the need to either share or aggregate massive amounts of data, FL can further democratize machine learning and scientific research.[10] FL is of particular interest in areas where the shared data are sensitive (operational data or personal information) or difficult (legal barriers to sharing).[11] Biomedical research involving personal

ABBREVIATIONS: AUROC, area under the ROC curve; **AWS,** Amazon Web Service; **CNS,** central nervous system; **CUDA,** Compute Unified Device Architecture; **FL,** federated learning; **GPU,** graphics processing unit; **ICH,** intracranial hemorrhage; **IT,** information technology; **RSNA,** Radiological Society of North America.

Supplemental digital content is available for this article at neurosurgery-online.com.

health information is a natural scenario for FL, and several recent publications have demonstrated the utility of this approach for predicting COVID-19 oxygen requirements,[12] analyzing mammography,[13] segmenting brain tumors,[14] and other clinical uses.[10]

Many published FL solutions have been the work of sophisticated computational teams, suggesting technical expertise likely limits widespread adoption and impact.[12,14-17] We hypothesized that with the right open-source software and infrastructure, disparate clinical teams can collaboratively deploy FL solutions to rapidly answer scientific questions of mutual interest. We assembled a multicenter research group of physician-scientists and trainees at 5 academic medical centers to assess the feasibility of this using the *nvflare* library, developed by NVIDIA (NVIDIA Corporation), to train a federated model on a prototypical neurosurgical problem: the classification of intracranial hemorrhage. While demonstrating the feasibility of using FL to train a predictive model on a clinical problem, our study reports on the challenges, pitfalls, and practical aspects of using FL in a clinical context while providing an open-source roadmap to assist other clinical teams in their collaborations.

METHODS

For the purposes of demonstrating the feasibility of a FL approach, we used a data set from the Radiological Society of North America (RSNA) hosted on the data science website Kaggle.[18] The data consist of two-dimensional slices of noncontrast head computed tomography scans that were annotated by radiologists with 6 possible labels: any bleed, epidural, intraparenchymal, intraventricular, subarachnoid, or subdural hemorrhage. These diagnoses are not mutually exclusive, and a single image can contain multiple labels (**Supplemental Digital Content 1, Figure 1**, http://links.lww.com/NEU/D422). Each patient was associated with (on average) approximately 40 images, and images were uniformly distributed across 5 participating centers on a per-patient basis to avoid potential data leakage across sites as follows: Site 1 (3409 patients), Site 2 (3409 patients), Site 3 (3409 patients), Site 4 (3408 patients), and Site 5 (3409 patients) (**Supplemental Digital Content 2, Figure 2**, http://links.lww.com/NEU/D423). A subset of the RSNA data (1894 patients, 74 937 images) was withheld for testing and benchmarking analyses. This study was exempt from institutional review and did not require patient consent as these data are publicly available and do not contain personal health information.

Raw Digital Imaging and Communications in Medicine files, a standard format for radiographic images, were organized at each site in a predetermined relative file path to ensure admin-deployed code had access to the correct data. Data were preprocessed at each site using a workflow known as "windowing," which adjusts the range of Hounsfield units displayed on a computed tomography image to enhance contrast between objects of varying densities. Each image was converted to a brain, subdural, and soft tissue window and stacked into an image containing pixel information from all 3 windows (Figure 1A). Processed data were then loaded into the pretrained convolutional neural network ResNeXt101 (32x8d) from the python-based PyTorch library Torchvision v. 0.11.1 (Meta Platforms, Inc.). Using standard machine learning techniques and model optimization, the neural network was trained on 85% of each sites' local data for 3 epochs and returned an output of predictions (Figure 1B). The loss function during local training was weighted based on each site's sample distribution to reflect imbalance in the data sets, where uncommon subtypes, such as epidural bleeds, were given a greater penalty if misdiagnosed by the model to reduce the rate of false-negatives. Through the *nvflare* workflow "ScatterAndGather," model weights from each site were returned to the central server for aggregation. After a round of training and central aggregation, the globally averaged model parameters were returned to each local client for another round of training. The process continued until model training was stopped (Figure 1C). After training, all local models and the global FL model underwent cross-site validation, during which each model was tested on 15% of each site's local data to evaluate model performance across the federated network.

Setup of the FL network was enabled by NVIDIA's domain-agnostic, open-source software development kit *nvflare*. The coordinating site (Site 2) used the *nvflare* provisioning tool to generate .zip files containing the code necessary to initialize the central server, admin, and each client (Figure 2). The central server and admin were hosted on an Amazon Web Service (AWS) EC2 instance (Amazon.com, Inc.; **Supplemental Digital Content 3, Table 1**, http://links.lww.com/NEU/D424). Each client .zip file was password-protected and emailed to the respective collaborator. On a local or cloud-based graphics processing unit (GPU), each client was responsible for installing Python 3, NVIDIA's Compute Unified Device Architecture (CUDA) toolkit and a virtual environment installed with *nvflare* and project-specific packages. After decompressing the file and running a simple start script using the command line, each client connected with the central server.

The global FL model was then compared with an identical neural network trained on the same data through centralized (nonfederated) training as a benchmark. The relative performance was compared using the python statistical package sklearn (v. 1.0.2) to calculate the area under the ROC curve (AUROC). We estimated the variance and CI associated with predictions for both cross-site validation and benchmarking using a binomial distribution with the R package Hmisc v. 4.7-0 (R Foundation). In addition, a qualitative survey was sent to all participating centers to obtain subjective feedback about their technical background and experience setting up the FL network. We wrote a playbook and an open-source codebase to enable the use of FL by practicing clinicians (**Supplemental Digital Content 4, Supplemental Playbook**, http://links.lww.com/NEU/D425).

RESULTS

Four of the 5 collaborating centers were able to successfully connect to the central server as a client. One site was unable to connect to the federated network in the timeframe of this study because of unforeseen IT and administrative lead time delays. In

(Continued from previous page)

*Department of Neurosurgery, NYU Langone Health, New York, New York, USA; ‡nVidia, Santa Clara, California, USA; §Center for Data Science, New York University, New York, New York, USA; ||Department of Neurosurgery, Vanderbilt University Medical Center, Nashville, Tennessee, USA; ¶Department of Neurosurgery, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania, USA; #Department of Neurosurgery, University of Michigan School of Medicine, Ann Arbor, Michigan, USA; **Department of Neurosurgery, Stanford University School of Medicine, Stanford, California, USA; ††Department of Ophthalmology, NYU Langone Health, New York, New York, USA; ‡‡Department of Population Health, NYU Langone Health, New York, New York, USA; §§Department of Radiology, NYU Langone Health, New York, New York, USA

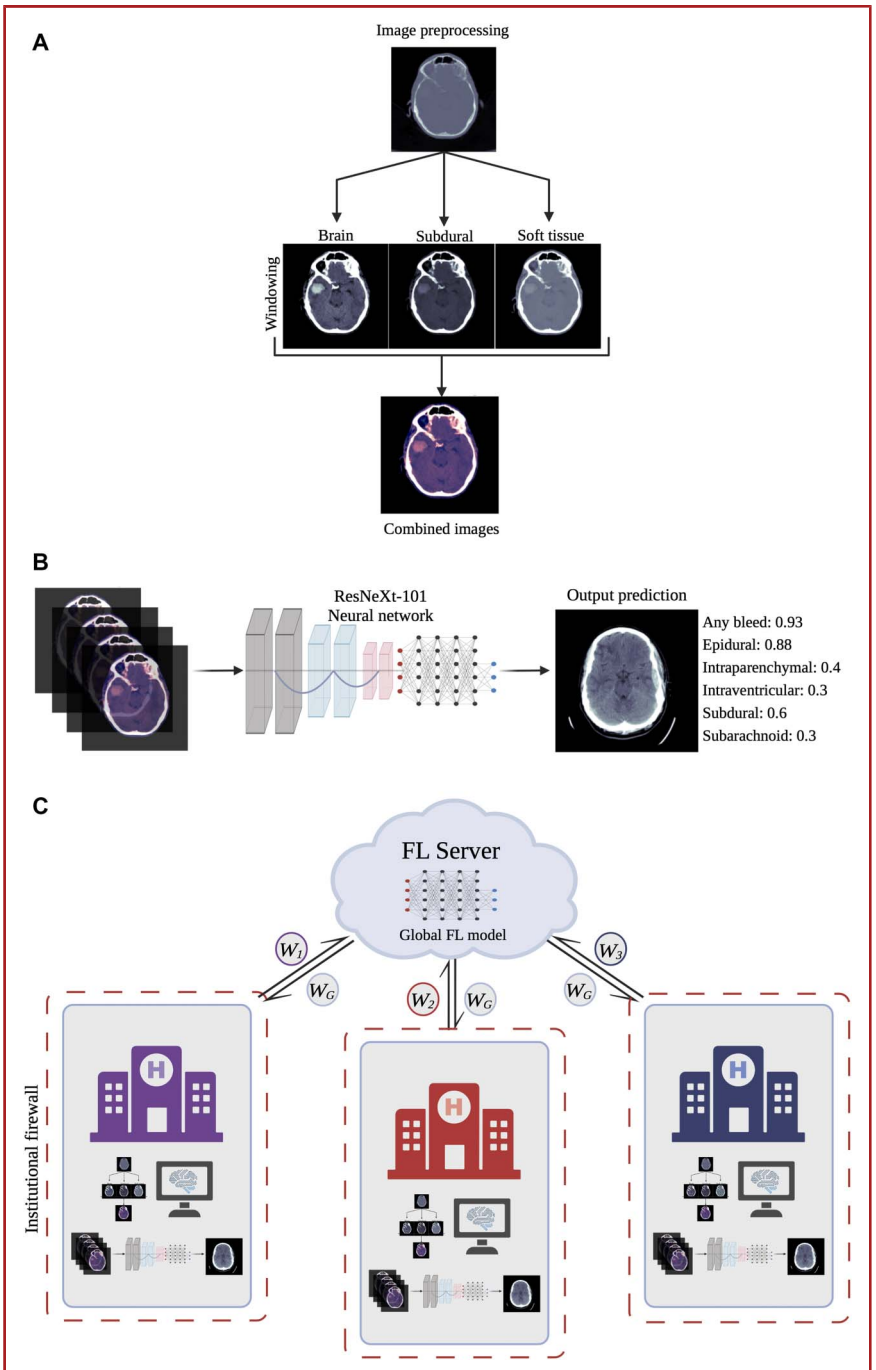

FIGURE 1. *Workflow of machine learning through a federated paradigm.* **A**, *Digital Imaging and Communications in Medicine file images were preprocessed using a workflow similar to that of a radiologist. Each image underwent 3 distinct windowing steps, which were then stacked together for training.* **B**, *Preprocessed slices were each passed through a pretrained ResNeXt101-32x8d convolutional neural network. The model then produced an independent prediction of the presence of each type of hemorrhage on a per slice basis.* **C**, *Federated learning works by training models on local sets of data and then transmitting the learned parameters to a central server for aggregation. After a round of training followed by central aggregation, the model parameters are sent back to the local clients for another round (K epochs) of training. This process continues until the model training is stopped. FL, federated learning.*

our experience, once a client had successfully connected, little to no additional human oversight or intervention from the client side was required. Cross-site validation revealed that although local models most often achieved the highest average AUROC when detecting all hemorrhage types on their own respective datasets, the global FL model achieved comparable and consistent performance when tested on data from any client (AUROC range: 0.935-0.9425) and was always one of the top three performing models regardless of the validation data's site of origin (Figure 3A; **Supplemental Digital Content 5, Table 2**, http://links.lww.com/NEU/D426). When tested on a separate data set withheld from all sites during training, the global FL model achieved a macroaverage AUROC of 0.9487 (95% CI 0.9471-0.9503) while the non-FL benchmark model achieved 0.9753 (95% CI 0.9742-0.9764; Figure 3B). FL global model performance varied by hemorrhage subtype, with subdural bleeds having the lowest AUROC (0.9257; 95% CI 0.9238-0.9276) and intraventricular having the highest (0.9751; 95% CI 0.9739-0.9762) on the holdout data set (**Supplemental Digital Content 6, Figure 3**, http://links.lww.com/NEU/D427).

The results from the qualitative survey revealed that of the 9 participating individuals involved in the practical setup of the federated network (n = 9; 100% response rate), 5 (56%) were practicing neurosurgeons with faculty positions, 1 (11%) was a neurosurgical resident, and 3 (33%) were medical students. Most participants (67%) had no formal education in computer science, and 33% had never heard of FL before this project (Table). Of the 4 sites able to connect to the federated network, 3 already had access to on-premises GPUs while 1 site required setting up a cloud-based GPU using AWS (**Supplemental Digital Content 3, Table 1**, http://links.lww.com/NEU/D424). GPU resources consisted of consumer-grade NVIDIA GPUs (RTX 2080 x2, RTX 3090 x1) at 4 sites and an enterprise grade GPU (V100) at 1 site in a cloud instance. Although connecting the *nvflare* client was uniformly described as "easy" by all 4 successfully connected sites and required less than an hour of total work to set up, the main reported difficulty occurred for the site requiring cloud-based computing because of university-specific administrative logistics. Only 1 client experienced unexpected disconnections from the FL server during the training process because of building generator power testing, which was resolved by connecting the GPU to an uninterruptible power supply battery.

DISCUSSION

In this study, we demonstrated the use of FL by a team of practicing clinicians and trainees to build a neurosurgical predictive model using open-source software. The accompaniment of this manuscript with a well-documented, open-source repository and FL "playbook" (**Supplemental Digital Content 4, Supplemental Playbook**, http://links.lww.com/NEU/D425) is unique and intended to accomplish several aims: (1) facilitate the reproducibility of this computational study, (2) provide a template for other groups to build upon and extend to other biomedical questions, and (3) encourage the development of an open-source community for physician-computer scientists to share their computational work.

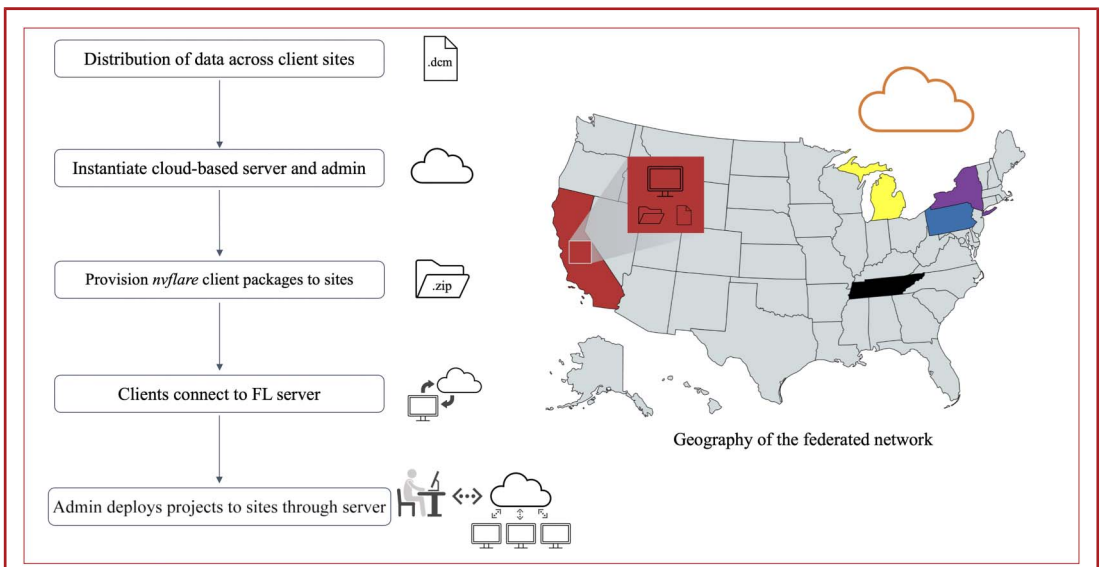

FIGURE 2. *We set out to establish a multi-institutional FL network with teams of clinicians across 5 participating sites. The lead site instantiated a cloud-based server for the federated network and dictated a FL schema by generating client packages for client site installation using the nvflare provisioning tool. Five sites spread across the United States installed the client packages to connect to the network server using these provisioned zip files, after which the admin could deploy projects and models for training. FL, federated learning.*

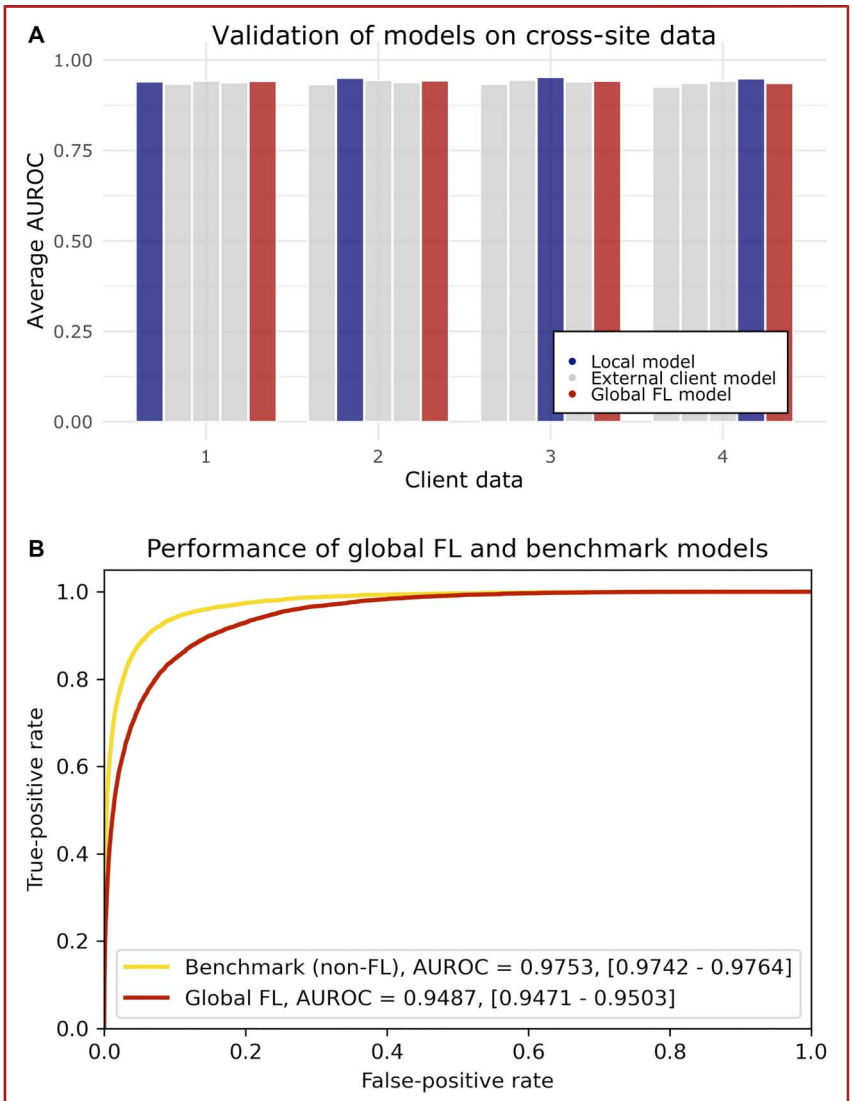

FIGURE 3. **A**, *Performance of local models and the global FL model in predicting all hemorrhage types when validated on data from each client site during cross-site validation. Local models validated on data from their own site (blue) performed expectedly well, although performance often diminished when the same model was tested on data from other sites (grey). The global FL model (red) achieved top 3 performance regardless of the data origin, suggesting improved generalizability across the federated network.* **B**, *Comparison of the global model and benchmark (standard, centrally trained) model on a hold-out data set for testing revealed comparable performance and high AUROC values, indicating the viability of an FL approach to intracranial hemorrhage detection. AUROC, area under the receiver operating characteristic curve; FL, federated learning.*

Although the global FL model achieved results comparable with that of the non-FL benchmark, the relatively lower AUROC is possibly due to the effects of training decentralized models using nonindependent and identically distributed (non-IID) data. This occurs when the various data sets used for FL training have inhomogeneous data distributions (because of factors such as varying imaging protocols or epidemiological differences) and the model has difficulty optimizing for both the global and site-specific solutions.[12] This is an active area of research, however, and novel strategies have been shown to mitigate the effects of non-IID data to harness FL's promise of generalizability.[10] Regardless, the primary goal of this study was to assess the feasibility of establishing a federated network by neurosurgeons for a problem relevant to neurosurgery, and our quantitative and

TABLE. Results from a Qualitative Survey to Elicit the Experience and General Sentiment from Participants in the Federated Learning Network

Major themes	Responses from qualitative survey (n = 9; 100% response rate)	
Position of participants	• Neurosurgical faculty, n = 5 (56%) • Neurosurgical resident, n = 1 (11%)	• Medical student, n = 3 (33%)
CS background	• Majority (67%) had no formal education in CS	• 33% of participants had never heard of FL before this project
General experience setting up the FL client	• >50% described setting up the client as "easy" • Most reported <1 h to set up the client	• Minimal work required after initial setup • Minimal troubleshooting required
Difficulties experienced	• Lack of GPU access • University-specific AWS account logistics and payment • COVID-related delays	• Concerns about funding and payment for cloud-based computing • University IT infrastructure access
Potential barriers	• Consistent formatting of data at each site • Funding for data storage and computing	• Up-front cost of initial GPU/client setup • Basic knowledge of the command line/Linux
Future FL applications in neurosurgery	• Including nonacademic centers and underrepresented patients in research • Rare brain tumor and disease research	• Multi-institutional outcomes (operative complications, length of stay, and costs) • Imaging-based diagnosis of CNS pathology

AWS, Amazon Web Service; CNS, central nervous system; CS, computer science; FL, federated learning; GPU, graphics processing unit; IT, information technology.

qualitative results suggest that FL is not only feasible but accessible for practicing surgeon scientists.

Although this project fundamentally involved the use of sophisticated areas of computer science, the software we used abstracted much of this away from the user to facilitate seamless collaboration despite minimal physician familiarity with the underlying codebase. In fact, our survey of sites revealed that user programming capabilities and familiarity with the Linux command line (a requisite for connecting the client) were not major barriers to implementation. This is a notable departure from prior studies that used highly technical teams of software engineers to train FL models.[12]

Even with accessible software, however, the process of building a multicenter federated network involved overcoming several challenges. The results from the qualitative survey showed that the major barriers to FL participation included access to GPUs, costs for storage and computing, and institutional administrative hurdles. Obtaining or maintaining access to GPU computing resources was the single largest challenge faced by sites and accounted for most setup time. Three of the successfully connected clients used on-premises hardware with relatively similar configurations. One site lacked on-premises GPU access but was able to rent a cloud-based GPU instance through an institutional AWS account. The FL server was also hosted on AWS because of the ease and flexibility for implementing the necessary networking protocols. One of our 5 member sites was not able to connect at all in the timeframe of this study because of university information technology lead time delays, which suggests that although federated networks allow rapid collaboration once established, FL is not immune to administrative delays and requires a relatively high investment upfront.

Despite these barriers, all participants reported high interest in pursuing FL-based projects in the future. Our qualitative survey found that topics pertinent to neurosurgery that could be facilitated by FL include improved access to rare brain tumor data; predictive modeling of operational considerations (such as length-of-stay) at a multi-institutional level and inclusion of a broader range of research collaborators—such as smaller institutions, nonacademic centers, and international collaborators—to improve generalizability of machine learning algorithms. As wearable medical devices become more commonplace and generate local data, FL could also be used to train a model while preserving the privacy of the wearer.[10]

Although we concede that the aforementioned projects could be addressed through a centralized machine learning paradigm, the traditional barriers to centralization of data nonetheless remain. This study demonstrates that FL can be implemented by a team of clinicians to provide a technical solution when traditional collaboration is neither feasible nor desired. FL can largely eliminate concerns for data breach when moving data off-site, and research questions that may be hindered by business or personal interests could benefit from a federated collaboration in which a single source's data cannot be exposed. For a field as small as neurosurgery, in particular, predictive modeling of surgical complications through FL could allow surgeons to openly and honestly share data regarding their cases while remaining confident that their personal outcomes remain anonymous. Moreover, studies using FL during the early phase of the COVID-19 pandemic demonstrated that this paradigm can facilitate rapid collaboration when faced with a novel emergency.[12] Neurosurgery has a long history of embracing new technologies, and FL at the very least offers another tool for conducting rapid and collaborative research.

Limitations

This study had several limitations. It was beyond the scope of this project to discuss the myriad of ways to configure a federated network with custom privacy and security features against malicious attack, which might be of interest to health care administrators and regulatory agencies.[10,19-21] In addition, although large and publicly available repositories such as the RSNA data set are critical for

machine learning prototyping, they are often modified for competition use.[22-24] They, therefore, do not present obstacles that may occur during FL on real-world data sets from disparate institutions and may require additional preprocessing to ensure conformity. As this proof-of-concept study aimed to determine the feasibility of setting up a federated network of clinicians, we did not exhaustively optimize the federated network hyperparameters to account for possible non-IID effects or weight the contribution of a client's model based on the relative size of the client's data set, although we attempted to reduce the potential impact of both by dividing the RSNA data set equally among all sites.[12,25,26] Finally, although *nvflare* significantly decreases the barrier-to-entry for client participation in FL, the coordinating center does require more extensive skills in computer science to orchestrate the experiments, and our project benefited from being located in a laboratory with access to an academic data science department.

CONCLUSION

Federated learning offers a privacy-preserving paradigm for multi-institutional collaboration and artificial intelligence research by effectively uniting silos of disparate health data. This study demonstrates the feasibility of establishing a federated network of practicing surgeon scientists using the open-source toolkit *nvflare* to address a prototypical neurosurgical emergency while also providing a roadmap for future clinicians. Our hope is that this study and accompanying codebase will lower the barrier-to-entry for FL projects that can address critical needs in health care in a timely, collaborative manner. Although FL promises to democratize machine learning research, we hope this project will democratize FL itself.

Funding

Supported in part by an Alpha Omega Alpha Carolyn L. Kuckein Student Research Fellowship (ATMC). This material is based upon work supported by the National Science Foundation under NSF Award 1922658. Dr Kwon, Dr Hollon, and Dr Aphinyanaphongs have funding from NIH.

Disclosures

Dr Yoon is a consultant for Johnson and Johnson, Bidermann Motech, and Pacira; has research grant support from Johnson and Johnson; if Founder of Kinesiometrics, INC, and MedCyclops, LLC. Dr Oermann's wife is a former employee of Merck, and a current employee of Mirati Therapeutics; Dr Oermann is a former employee of Google Inc.

REFERENCES

1. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. *Nat Med.* 2019;25(1):44-56.
2. Sun C, Shrivastava A, Singh S, Gupta A. Revisiting unreasonable effectiveness of data in deep learning era. arXiv [csCV]. Published online July 10, 2017. http://arxiv.org/abs/1707.02968
3. Asher AL, McCormick PC, Selden NR, Ghogawala Z, McGirt MJ. The national neurosurgery quality and outcomes database and NeuroPoint Alliance: rationale, development, and implementation. *Neurosurg Focus.* 2013;34(1):E2.
4. Gaspar L, Scott C, Rotman M, et al. Recursive partitioning analysis (RPA) of prognostic factors in three Radiation Therapy Oncology Group (RTOG) brain metastases trials. *Int J Radiat Oncol Biol Phys.* 1997;37(4):745-751.
5. Bydon M, Schirmer CM, Oermann EK, et al. Big Data defined: a practical review for neurosurgeons. *World Neurosurg.* 2020;133:e842-e849.
6. Titano JJ, Badgeley M, Schefflein J, et al. Automated deep-neural-network surveillance of cranial images for acute neurologic events. *Nat Med.* 2018;24(9):1337-1341.
7. Zech J, Pain M, Titano J, et al. Natural language-based machine learning models for the annotation of clinical radiology reports. *Radiology.* 2018;287(2):570-580.
8. Oermann EK, Kress MAS, Collins BT, et al. Predicting survival in patients with brain metastases treated with radiosurgery using artificial neural networks. *Neurosurgery.* 2013;72(6):944-951.
9. Brendan McMahan H, Moore E, Ramage D, Hampson S, Arcas BAY. Communication-efficient learning of deep networks from decentralized data. arXiv [csLG]. Published online February 17, 2016. http://arxiv.org/abs/1602.05629
10. Rieke N, Hancox J, Li W, et al. The future of digital health with federated learning. *NPJ Digit Med.* 2020;3:119.
11. van Panhuis WG, Paul P, Emerson C, et al. A systematic review of barriers to data sharing in public health. *BMC Public Health.* 2014;14(1):1144.
12. Dayan I, Roth HR, Zhong A, et al. Federated learning for predicting clinical outcomes in patients with COVID-19. *Nat Med.* 2021;27(10):1735-1743.
13. Roth HR, Chang K, Singh P, et al. Federated learning for breast density classification: a real-world implementation. In: Albarqouni S, et al., eds. *Domain Adaptation and Representation Transfer, and Distributed and Collaborative Learning*. Springer; 2020:181-191.
14. Pati S, Baid U, Zenk M, Edwards B, Sheller M. *The Federated Tumor Segmentation (FeTS) Challenge*. arXiv. Published online 2021. https://arxiv.org/abs/2105.05874
15. Kaissis G, Ziller A, Passerat-Palmbach J, et al. End-to-end privacy preserving deep learning on multi-institutional medical imaging. *Nat Mach Intell.* 2021;3(6):473-484.
16. Sui D, Chen Y, Zhao J, Jia Y, Xie Y, Sun W. Feded: federated learning via ensemble distillation for medical relation extraction. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), aclweb.org*; 2020:2118-2128. https://www.aclweb.org/anthology/2020.emnlp-main.165.pdf
17. Warnat-Herresthal S, Schultze H, Shastry KL, et al. Swarm learning for decentralized and confidential clinical machine learning. *Nature.* 2021;594(7862):265-270.
18. RSNA intracranial hemorrhage detection. Kaggle. Accessed December 30, 2022. https://www.kaggle.com/c/rsna-intracranial-hemorrhage-detection
19. Tolpegin V, Truex S, Gursoy ME, Liu L. Data poisoning attacks against federated learning systems. arXiv [csLG]. Published online July 16, 2020. http://arxiv.org/abs/2007.08432
20. Sun G, Cong Y, Dong J, Wang Q, Liu J. Data poisoning attacks on federated machine learning. arXiv [csCR]. Published online April 19, 2020. http://arxiv.org/abs/2004.10020
21. Fung C, Yoon CJM, Beschastnikh I. Mitigating sybils in federated learning poisoning. arXiv [csLG]. Published online August 14, 2018. http://arxiv.org/abs/1808.04866
22. Flanders AE, Prevedello LM, Shih G, et al. Construction of a machine learning dataset through collaboration: the RSNA 2019 brain CT hemorrhage challenge. *Radiol Artif Intell.* 2020;2(3):e190211.
23. Bakas S, Reyes M, Jakab A, Bauer S. Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the BRATS challenge. arXiv. Published online 2018. http://arxiv.org/abs/1811.02629
24. Russakovsky O, Deng J, Su H, et al. ImageNet Large SCALE visual recognition challenge. arXiv [csCV]. Published online September 1, 2014.
25. Hsieh K, Phanishayee A, Mutlu O, Gibbons PB. The non-IID data quagmire of decentralized machine learning. arXiv [csLG]. Published online October 1, 2019.
26. Xu J, Glicksberg BS, Su C, et al. Federated learning for healthcare informatics. *J Healthc Inform Res.* 2021;5:1-19.

Acknowledgments

We would like to thank our friends and collaborators at nVidia, Anthony Costa, PhD, and Mona Flores, MD, for their conversations, inspiration, and discussion around the use of federated learning in medicine and the constant need to democratize machine learning.

Supplemental digital content *is available for this article at neurosurgery-online.com.*

Supplemental Figure 1. Examples of preprocessing and windowing of computed tomography (CT) scans with ground-truth labels and diagnoses from radiologists.

Supplemental Figure 2. Data distribution across sites by hemorrhage subtype.

Supplemental Figure 3. Global federated learning (FL) and benchmark model area under the receiver operating characteristic curve (AUROC) performance for each hemorrhage subtype.

Supplemental Table 1. Hardware and graphics processing unit (GPU) specifications used by each participating member of the federated network.

Supplemental Table 2. Results from cross-site validation, where each local model and the global federated learning (FL) model are tested on a withheld subset of each site's local data. Average area under the ROC curve (AUROC) and 95% confidence intervals are reported.

Supplemental Playbook. Federated Learning in Health care: A roadmap and resource for clinicians. We have created a "playbook" that outlines the steps necessary to recreate findings experiment and assist others in using *nvflare* to set up a federated network. It provides the basic hardware, technical expertise, and associated screenshots of the setup to allow easy reproducibility and aid in troubleshooting.