



Federated learning: a collaborative effort to achieve better medical imaging models for individual sites that have small labelled datasets

Dianwen Ng¹, Xiang Lan¹, Melissa Min-Szu Yao^{2,3,4^}, Wing P. Chan^{2,3,4^}, Mengling Feng¹

¹Saw Swee Hock School of Public Health, National University Health System, National University of Singapore, Singapore; ²Department of Radiology, Wan Fang Hospital, Taipei Medical University, Taipei; ³Department of Radiology, School of Medicine, College of Medicine, Taipei Medical University, Taipei; ⁴Medical Innovation Development Center, Wan Fang Hospital, Taipei Medical University, Taipei

Correspondence to: Wing P. Chan, MD. Department of Radiology, Wan Fang Hospital, Taipei Medical University, 111 Hsing Long Road, Section 3, Taipei. Email: wingchan@tmu.edu.tw; Mengling Feng, PhD. Saw Swee Hock School of Public Health, National University Health System, National University of Singapore, Tahir Foundation Building 12 Science Drive 2, #10-01, Singapore. Email: ephfm@nus.edu.sg.

Abstract: Despite the overall success of using artificial intelligence (AI) to assist radiologists in performing computer-aided patient diagnosis, it remains challenging to build good models with small datasets at individual sites. Because many medical images do not come with proper labelling for training, this requires radiologists to perform strenuous labelling work and to prepare the dataset for training. Placing such demands on radiologists is unsustainable, given the ever-increasing number of medical images taken each year. We propose an alternative solution using a relatively new learning framework. This framework, called federated learning, allows individual sites to train a global model in a collaborative effort. Federated learning involves aggregating training results from multiple sites to create a global model without directly sharing datasets. This ensures that patient privacy is maintained across sites. Furthermore, the added supervision obtained from the results of partnering sites improves the global model's overall detection abilities. This alleviates the issue of insufficient supervision when training AI models with small datasets. Lastly, we also address the major challenges of adopting federated learning.

Keywords: Artificial intelligence (AI); data; federated learning; medical imaging

Submitted Apr 21, 2020. Accepted for publication Sep 25, 2020.

doi: 10.21037/qims-20-595

View this article at: <http://dx.doi.org/10.21037/qims-20-595>

Motivation

Artificial intelligence (AI) has proven successful for the detection and diagnosis of pathological conditions on medical images. Images, such as radiographs and computed tomography scans in Digital Imaging and Communications in Medicine (DICOM) format, are used by AI to self-learn and identify patterns that allow for the derivation of reasonable predictions. Even without the input of standard image-processing rules by medical experts,

network algorithms can learn from raw images, and ultimately provide highly accurate and consistent output. Furthermore, AI has demonstrated the potential to assist radiologists in performing computer-aided analysis and diagnosis. For example, Stanford University published a convolutional neural network with 121 layers, CheXNet, that outperformed four radiologists after learning on a dataset of 112,120 frontal-view chest radiographs (Chest X-ray14) (1). Likewise, Li *et al.* (2) presented a retrospective

[^] ORCID: Melissa Min-Szu Yao, 0000-0002-1295-4771; Wing P. Chan, 0000-0001-8322-6336.

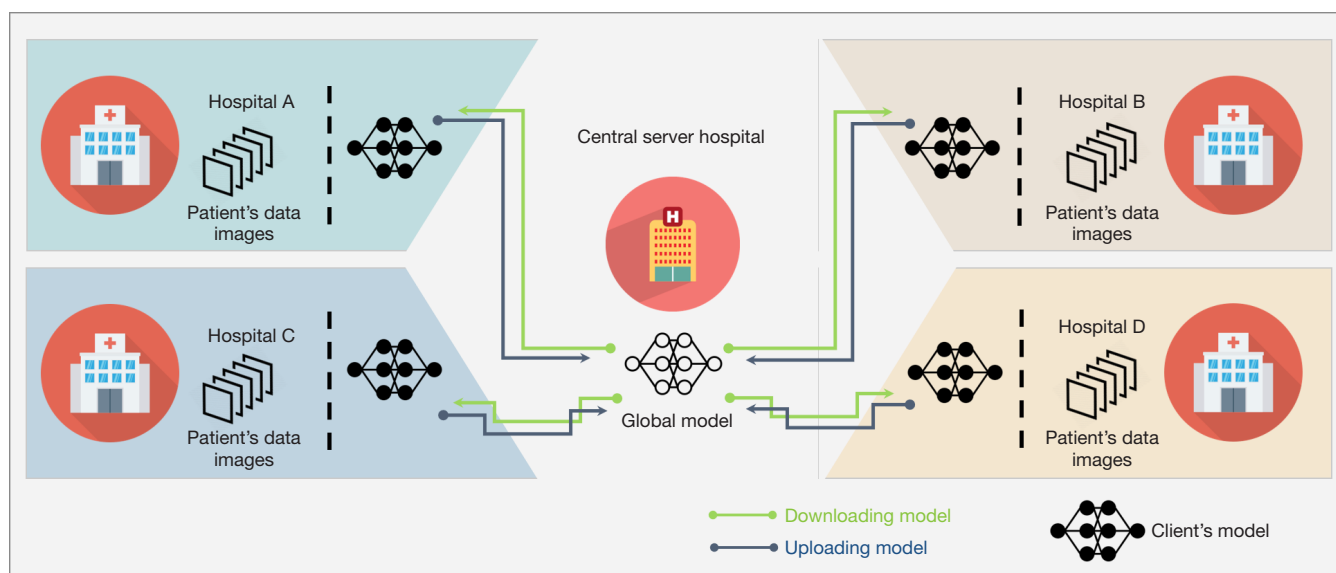


Figure 1 An overview of federated learning.

pre-clinical study that used deep convolutional neural networks to improve thyroid cancer diagnostic accuracy on sonographic images; the proposed network model attained a similar accuracy to the average performance of six radiologists but with a higher sensitivity.

Nonetheless, the high-accuracy performance model has a shortcoming: deep-learning solutions notoriously require much training data. Although many imaging centres own large imaging databases, many of the images are unlabelled, and therefore do not allow the model to learn. Unlabelled images must be manually annotated by a trained radiologist. Thus, producing a training dataset can be time-consuming and is a significant limiting factor in the development of AI solutions for medical imaging. To solve this problem, federated learning trains algorithms across multiple healthcare institutions to achieve better AI models through collaboration. A given algorithm gathers labelled patient information from various institutions to augment its learning base and therefore augment its ability to detect patients from a wide population. However, the idea of collaborative training has triggered concern among some healthcare administrators because of the possibility that personal information will be shared, and therefore breach the Data Protection Act and endanger patient privacy. To combat these concerns, federated learning allows an algorithm to learn from local data without removing that data from an individual site, thus upholding patients' privacy and facilitating cooperation between hospitals.

Overview of federated learning

Through federated learning, multiple organizations or institutions work together to solve a machine-learning problem under the coordination of a central server or service provider. Thus, a deep-learning model is maintained and improved upon within a central server. The model is trained by distributing itself to remote silo data centres, such as hospitals or other medical institutions, which allows these sites to keep their data localized. Data from each collaborator is never exchanged or transferred during training. Instead of bringing the data to the central server, as in conventional deep learning, the central server maintains a global shared model, which is disseminated to all institutions. Each entity subsequently maintains a separate model based on its own patients' data. Thereafter, each centre provides feedback to the server based on its individually trained model—either by its weight or the error gradient of the model. The central server aggregates the feedback from all participants, and based on predefined criteria, updates the global model. The predefined criteria allow the model to evaluate the quality of the feedback and therefore to only incorporate that which is value-adding. The feedback from centres with adverse or strange results can thus be ignored. This process forms one round of federated learning, and it is iterated until the global model is trained.

Figure 1 illustrates the federated learning process, and

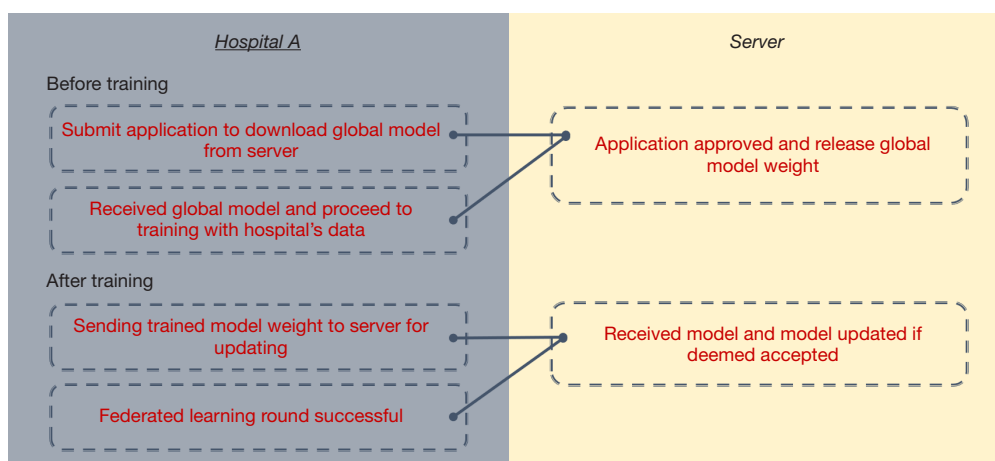


Figure 2 Summary of one complete round of federated learning.

Figure 2 summarizes the learning framework procedures. Federated learning allows individual hospitals to benefit from the rich datasets of multiple non-affiliated hospitals without centralizing the data in one place. This practice overcomes critical issues such as data privacy, data security, data access rights, and access to heterogeneous data. Hence, federated learning allows multiple collaborators to build a robust machine-learning model using a large dataset.

The benefits of federated learning

First, federated learning allows the central model to learn from a diverse and augmented set of learning samples obtained from multiple institutions. Patient data and images at any individual hospital is obtained from a specific subset population and is therefore unlikely to have been seen by or shared with other hospitals. This is particularly true of hospitals located in different geographic regions, where patient traits likely differ substantially. The gender ratios, age distributions, and ethnicities of the patient populations all tend to differ between hospitals. Likewise, a tertiary hospital tends to see a higher volume of difficult cases than a secondary hospital does. Because of these differences in patient populations, if an atypical patient is evaluated at a certain hospital, the use of a training model derived from that hospital's patient database may be inadvisable, as the model has not had the opportunity to learn from atypical cases. However, a federated learning derived model incorporates data from multiple institutions, thereby increasing its external validity. Such a model would be much more likely to generate accurate results, even for

what may be an atypical patient at a certain hospital. This type of cooperation therefore allows for the advancement of precision medicine.

Second, the deployment of AI models requires periodic training and updating to remain current. This requirement may place an undue burden on radiologists, who must continually label a sufficient volume of studies necessary to retrain the model. At certain times when patient volume peaks, it may be difficult or even impossible for radiologists to produce enough annotated labels. However, because peak volume season may differ across hospitals, federated learning mitigates this issue, allowing radiologists at less busy hospitals to annotate studies while their counterparts at busier hospitals are too busy to do so. In this way, all users can download and use the most up to date model year round.

Third, the federated learning framework brings about auto-scaling at almost no additional cost. When new hospitals participate, they bring more data and more computational resources. As the loop continues to run, an ever-enlarging dataset is fed to the model, while all computations continue to be made by the end-user. The global model is updated after users have trained their individual models, requiring minimal resources to aggregate models and thus making deployment much more economical.

Application of federated learning

The concept of federated learning is a new and popular research topic and is being widely explored in healthcare. Numerous reports have demonstrated proof of concept with respect to federated learning applied to real-world

medical imaging. In 2018, Intel worked with the Centre for Biomedical Image Computing and Analytics, at the University of Pennsylvania, to evaluate the use of federated learning for brain-image segmentation (3). The publicly available dataset from the brain tumour segmentation (BraTS) challenge 2018 was used (4-7). The data were a collection of multi-institutional, multi-model magnetic resonance imaging (MRI) brain scans from patients with gliomas. Each of the abnormal findings on the MRI scans was manually annotated by as many as four radiologists using three distinct labels that corresponded to either peritumoral edematous/invaded tissue, non-enhancing/solid, necrotic/cystic tumour core, or enhancing tumour. The authors chose the U-Net deep convolutional neural networks model for the task, while deploying the server-client federated learning algorithm for the system environment to train and perform model validation. Multiple hypothetical institutions were created to simulate independent, separate clients. Data were subsequently split and tested in two ways. The first split evenly distributed and randomly allocated data to each client; the second split assigned the data to the institutions from which they were initially collected. During implementation, multiple clients received the current version of the model from the central server. The server selected a few suitable models from individual clients and updated the central model using federated aggregation. This update, in theory, allows the central model to increase performance and accuracy, thereby allowing clients that receive the updated model from the server in the next round to perform better. In the end, the BraTS experiments revealed that the performance scores of the federated semantic segmentation models on the brain MRI scans were similar to those derived from models that trained on the complete dataset.

In similar research, Nvidia Corporation worked with King's College London and presented their work at the Medical Image Computing and Computer Assisted Intervention Society (MICCAI 2019) conference. Federated learning training was performed on Nvidia Clara Train SDK (8). Using the BraTS 2018 dataset, they attempted to apply the differential-privacy technique to protect patient data under a federated learning setup. This technique encodes the data of each patient before sharing the information with other clients. Complex mathematical algorithms are employed to prevent reverse engineering and restoration of the original dataset. Ultimately, Nvidia was able to achieve comparable segmentation performance using the federated learning model without directly sharing

institutional data.

In the above experiments, clients were endowed with a small dataset that simulated real-life healthcare scenarios, where there is never enough labelled data. However, with federated learning, members of the participating community can obtain performance that is akin to training with a large dataset. Therefore, federated learning offers a way to bypass the problem of not having sufficient labelled data to deploy top-level machine-learning solutions, through the combined effort of various medical institutions. Thus, a well-performing model can be generated that offers strong external validity based on large datasets without worrying about data protection.

Challenges of federated learning

The first challenge of performing federated learning successfully is weights updating. This challenge is faced during the training phase, when the deep-learning model for medical-imaging analysis uses backpropagation for optimization. In federated learning, the central server must aggregate the weights from various hospitals during backpropagation, and the model-aggregation policy can directly affect model performance: an efficient and high-performance strategy for the updating of weights is crucial for the implementation of federated learning.

A potential solution for addressing weights updating is to use federated averaging, as proposed by McMahan *et al.* (9). Using this strategy, the central server incorporates the average weights of all selected models following each iteration to create the central model. As such, all site-specific models are weighted equally, regardless of the fact that some are likely superior to others. Weaker site-specific models may result from poor data quality, lack of target patients, or annotation errors made by an inexperienced radiologist. Thus, feedback from some hospitals should ideally be weighted less than feedback from other hospitals. Federated averaging can in this way obscure the effects of important features and is therefore not an ideal solution to weighting. A suitable solution to the problem of weighting is a persistent challenge that deserves further study.

The second challenge of performing federated learning is the equitable allocation of grant funding. When multiple hospitals collaborate, larger hospitals tend to generate more images, have more-experienced radiologists doing the labelling, and have better training infrastructure. Ideally, these collaborators contribute more to joint learning and produce a higher quality dataset for better training

feedback. However, doing so is costly, so it is natural for larger hospitals to expect greater research-grant funding. Some may argue though, that research funds should be apportioned according to the value of training feedback, with top value-adding contributors receiving more. A problem with this approach is that determining the value of contributions is difficult and should not be solely based on the size of datasets. For example, a hospital may produce ten times more images than another, but the diversity of its images may be low. In such a scenario, it offers little value in training the model to recognize various pathologic conditions. Furthermore, it is unclear how best to qualify the judgement of radiologists when labelling images, especially for semantic segmentation. Hence, algorithms are needed to better appraise the contributions made by individual hospitals and thus to more objectively allocate funding.

The third challenge lies in the practical application of federated learning. Hardware, operating systems, and network conditions differ across sites, which means that learning algorithms must run on different platforms. This poses a challenge to the implementation of federated learning. For example, graphics processing units may differ across hospitals, which results in differences in speed of training and asynchronous weights updating. Moreover, since algorithms are running remotely, central data scientists have no direct control over or may not understand individual institution specifics. This can make optimization and debugging difficult and complex.

The fourth challenge arises from differences in image acquisition protocols and labelling methodologies across institutions. Such differences may lead to the generation of a site specific model that does not fit other sites well and therefore may contribute poorly or even negatively to the central model. Such differences may also hold important implications in weighting across sites. This can be overcome by proactively agreeing upon and implementing certain processing and labeling standards across all involved institutions. The difficulties of achieving such standardization can be reduced by using the same natural language processing (NLP) algorithms to process radiologist reports. This is the current practice adopted by many institutions in preparing their data for training models.

Conclusions

To conclude, it is extremely difficult for individual sites that

have small labelled datasets to build their own AI models for patient diagnosis. This is because the scope of patients that the model is based upon is typically quite limited, which in turn considerably limits the external validity of the model. With federated learning, this barrier is eliminated. Through collaboration, multiple institutions can pool their data to train a global model that offers greater accuracy over a larger spectrum of patients. In this collaborative effort, there is no direct data sharing, as federated learning prioritizes the privacy of patient data. Instead, the process of federated aggregating permits the generation of a central model based on recurrent updates from individual sites. Federated learning offers easy scalability, flexible training scheduling, and large training datasets through multi-site collaborations, all essential conditions to the successful deployment of an AI solution. However, important challenges remain and must be addressed before federated learning is optimally able to build AI models. Further, because of the novelty of federated learning in medical imaging AI, this topic has the potential to inspire and attract researchers, whose work will be necessary to advance the field forward.

Acknowledgments

Funding: This research is supported by the National Research Foundation Singapore under its AI Singapore Programme (Award No. AISG-GC-2019-002), the NUHS Joint Grant (WBS R-608-000-199-733) and the NMRC Health Service Research Grant (HSRG-OC17nov004).

Footnote

Conflicts of Interest: All authors have completed the ICMJE uniform disclosure form (available at <http://dx.doi.org/10.21037/qims-20-595>). The authors have no conflicts of interest to declare.

Open Access Statement: This is an Open Access article distributed in accordance with the Creative Commons Attribution-NonCommercial-NoDerivs 4.0 International License (CC BY-NC-ND 4.0), which permits the non-commercial replication and distribution of the article with the strict proviso that no changes or edits are made and the original work is properly cited (including links to both the formal publication through the relevant DOI and the license). See: <https://creativecommons.org/licenses/by-nc-nd/4.0/>.

References

1. Rajpurkar P, Irvin J, Zhu K, Yang B, Mehta H, Duan T, et al. Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. arXiv preprint arXiv:1711.05225. 2017. Available online: <https://arxiv.org/abs/1711.05225>
2. Bakas S, Akbari H, Sotiras A, Bilello M, Rozycki M, Kirby JS, Freymann JB, Farahani K, Davatzikos C. Advancing The Cancer Genome Atlas glioma MRI collections with expert segmentation labels and radiomic features. *Sci Data*. 2017;4:170117.
3. Bakas S, Akbari H, Sotiras A, Bilello M, Rozycki M, Kirby J, Freymann J, Farahani K, Davatzikos C. Segmentation labels and radiomic features for the pre-operative scans of the TCGA-GBM collection. Available online: <https://wiki.cancerimagingarchive.net/display/DOI/Segmentation+Labels+and+Radiomic+Features+for+the+Pre-operative+Scans+of+the+TCGA-GBM+collection#242826662c5ce8901dc84f4393fdccced7375a3c>
4. Bakas S, Akbari H, Sotiras A, Bilello M, Rozycki M, Kirby J, Freymann J, Farahani K, Davatzikos C. Segmentation labels and radiomic features for the pre-operative scans of the TCGA-LGG collection. Available online: <https://wiki.cancerimagingarchive.net/display/DOI/Segmentation+Labels+and+Radiomic+Features+for+the+Pre-operative+Scans+of+the+TCGA-LGG+collection>
5. Li W, Milletari F, Xu D, Rieke N, Hancox J, Zhu W. Privacy-preserving federated brain tumour segmentation. *International Workshop on Machine Learning in Medical Imaging*. Springer, 2019.
6. Konečný J, McMahan HB, Yu FX, Richtárik P, Suresh AT, Bacon D. Federated learning: Strategies for improving communication efficiency. arXiv preprint arXiv:1610.05492. 2016. Available online: <https://arxiv.org/abs/1610.05492>
7. Li X, Zhang S, Zhang Q, Wei X, Pan Y, Zhao J, Xin X, Qin C, Wang X, Li J, Yang F, Zhao Y, Yang M, Wang Q, Zheng Z, Zheng X, Yang X, Whitlow CT, Gurcan MN, Zhang L, Wang X, Pasche BC, Gao M, Zhang W, Chen K. Diagnosis of thyroid cancer using deep convolutional neural network models applied to sonographic images: a retrospective, multicohort, diagnostic study. *Lancet Oncol* 2019;20:193-201.
8. Sheller MJ, Reina GA, Edwards B, Martin J, Bakas S (editors). Multi-institutional deep learning modeling without sharing patient data: A feasibility study on brain tumor segmentation. *International MICCAI Brainlesion Workshop*. Springer, 2018.
9. Menze BH, Jakab A, Bauer S, Kalpathy-Cramer J, Farahani K, Kirby J, Burren Y, Porz N, Slotboom J, Wiest R, Lanczi L, Gerstner E, Weber MA, Arbel T, Avants BB, Ayache N, Buendia P, Collins DL, Cordier N, Corso JJ, Criminisi A, Das T, Delingette H, Demiralp Ç, Durst CR, Dojat M, Doyle S, Festa J, Forbes F, Geremia E, Glocker B, Golland P, Guo X, Hamamci A, Iftekharuddin KM, Jena R, John NM, Konukoglu E, Lashkari D, Mariz JA, Meier R, Pereira S, Precup D, Price SJ, Raviv TR, Reza SM, Ryan M, Sarikaya D, Schwartz L, Shin HC, Shotton J, Silva CA, Sousa N, Subbanna NK, Szekely G, Taylor TJ, Thomas OM, Tustison NJ, Unal G, Vasseur F, Wintermark M, Ye DH, Zhao L, Zhao B, Zikic D, Prastawa M, Reyes M, Van Leemput K. The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS). *IEEE Trans Med Imaging* 2015;34:1993-2024.

Cite this article as: Ng D, Lan X, Yao MMS, Chan WP, Feng M. Federated learning: a collaborative effort to achieve better medical imaging models for individual sites that have small labelled datasets. *Quant Imaging Med Surg* 2021;11(2):852-857. doi: 10.21037/qims-20-595