Collaborative Privacy-preserving Approaches for Distributed Deep Learning Using Multi-Institutional Data

Sharut Gupta, BTech, MTech • Sourav Kumar, BTech* • Ken Chang, MD, PhD • Charles Lu, BS*
Praveer Singh, PhD • Jayashree Kalpathy-Cramer, PhD

** S.G. and S.K. contributed equally to this work.*
Author affiliations, funding, and conflicts of interest are listed at the end of this article.

Deep learning (DL) algorithms have shown remarkable potential in automating various tasks in medical imaging and radiologic reporting. However, models trained on low quantities of data or only using data from a single institution often are not generalizable to other institutions, which may have different patient demographics or data acquisition characteristics. Therefore, training DL algorithms using data from multiple institutions is crucial to improving the robustness and generalizability of clinically useful DL models. In the context of medical data, simply pooling data from each institution to a central location to train a model poses several issues such as increased risk to patient privacy, increased costs for data storage and transfer, and regulatory challenges. These challenges of centrally hosting data have motivated the development of distributed machine learning techniques and frameworks for collaborative learning that facilitate the training of DL models without the need to explicitly share private medical data. The authors describe several popular methods for collaborative training and review the main considerations for deploying these models. They also highlight publicly available software frameworks for federated learning and showcase several real-world examples of collaborative learning. The authors conclude by discussing some key challenges and future research directions for distributed DL. They aim to introduce clinicians to the benefits, limitations, and risks of using distributed DL for the development of medical artificial intelligence algorithms.

Introduction

Deep learning (DL), a subfield of machine learning, has demonstrated remarkable capability for automation of various tasks in medical imaging analysis and radiologic reporting (1). Despite the applicability of DL tools to many potential clinical workflows, one major challenge is that they require large, diverse datasets to train robust models that can be deployed in clinical practice. Currently, many DL models are trained only on a single institution's internal data resources, limiting the quantity and diversity of patient populations, data acquisition characteristics, and abnormalities represented in the dataset that are available to develop generalizable DL tools (Fig 1A). Even when multi-institutional data cohorts are used to train DL algorithms, the dataset may still disproportionately underrepresent certain patient demographics or geographic regions (2). Training on insufficiently diverse datasets may result in DL models that are brittle throughout different settings, time periods, or institutions (3,4). This limitation is often overlooked in published studies (5) in which performance evaluation is not reported on external testing datasets. Furthermore, training

Supplemental Material

RadioGraphics 2023; 43(4):e220107
https://doi.org/10.1148/rg.220107
Content Code: IN

CME

Quiz questions for this article are available in the supplemental material.

Abbreviations: DL = deep learning, IID = independent and identically distributed

TEACHING POINTS

- Even when multi-institutional data cohorts are used to train DL algorithms, the dataset may still disproportionately underrepresent certain patient demographics or geographic regions. Training on insufficiently diverse datasets may result in DL models that are brittle throughout different settings, time periods, or institutions.
- Currently, there are three common approaches to developing DL models using multi-institutional data: training on centrally hosted data, pretraining on public domain datasets, and transfer learning.
- The inadequacies of current approaches to the multi-institutional collaboration of DL algorithms for clinical applications have motivated the development of new methodological frameworks for distributed collaborative learning. These techniques allow collaborations to benefit from training DL models of their collective datasets without explicitly sharing the data.
- A popular approach for federated learning is federated averaging, which involves training a local copy of the model at each institution that is then aggregated at a central server, usually by averaging the weights of all the local models to form a global model.
- Distribution shift is a major challenge for distributed learning, because collaborative learning among institutions with different data distributions can negatively affect performance.

on datasets with underrepresented demographics or attributes can lead to biased models that, when deployed, may reinforce and propagate existing disparities among vulnerable and disadvantaged groups in health care and in society. Access to larger quantities of diverse training data should be prioritized to ensure that advances in artificial intelligence technologies such as DL are maximally generalizable, without harmful biases, to be suitable for clinical deployment (6).

Currently, there are three common approaches to developing DL models using multi-institutional data: training on centrally hosted data, pretraining on public domain datasets, and transfer learning. The first approach is to centrally host the data (Fig 1B). This involves transferring data from each institution to a single location where a DL algorithm can be trained using the aggregated data. Although it is arguably the most straightforward approach, centrally hosting data creates several privacy concerns and logistic challenges. First, regulatory issues regarding the negligent sharing of private medical data may lead to points of contention with the Health Insurance Portability and Accountability Act (HIPAA) and General Data Protection Regulation (GDPR) in Europe. In addition, repeatedly transferring and storing large quantities of data from each institution comes with server costs and security risks for data breaches.

The second approach to obtaining access to large quantities of data for multi-institutional collaborations is through datasets available in the public domain. Although this allows easy access to data resources for existing and new collaborations, there are additional complications in making data publicly available. A major challenge is the necessity of anonymizing protected health information, which can be a difficult and time-consuming process because radiologic reports and Digital Imaging and Communications in Medicine (DICOM) metadata must be stripped of any private or sensitive information. Although there have been several successful examples of medical datasets being made public, such as those of the National Lung Screening Trial (7), the Medical Imaging Data Resource Center (8), the Brain Tumor Image Segmentation (BRATS) (9) challenge, and the Alzheimer Disease Neuroimaging Initiative (10), collecting and curating medical datasets can be prohibitively expensive. In addition, public medical datasets may have the possibility of unintended usage (eg, commercialization or nefarious purposes), which may discourage institutions from releasing their data. Furthermore, the misalignment of the interests of hospitals and investigators to maintain a competitive advantage in the research and/or clinical arena hinders the release of publicly available datasets. Finally, proprietary datasets are increasingly viewed as valuable resources because algorithms can be considered intellectual property that may disincentivize openly sharing large medical datasets.

A third approach is to take a model that was already trained on data at one institution and perform subsequent training on data from another institution, which is a procedure called *transfer learning*. However, this approach can result in a model that is overfitted to specific characteristics of a single institution's data at the cost of lower generalizability to data distributions of other external institutions. Thus, the overly fine-tuned algorithm may need to be retrained for every new institution. The other challenge with this approach is that continuous training of algorithms on new datasets may result in a decrease in prediction accuracy on data from previously trained institutions, a phenomenon known as *catastrophic forgetting* (11). Although addressing catastrophic forgetting is an active area of research (12), this degradation in the model's performance as a result of transfer learning poses a major challenge to the reliability of the algorithm and complicates risk management for regulatory agencies, such as the Food and Drug Administration (FDA) in the United States, because models that have been fine-tuned after their approval may no longer satisfy the required performance criteria on the original test set. These complications in the regulation of DL-based medical device software may further discourage research and development among multi-institutional collaborations.

The inadequacies of current approaches to the multi-institutional collaboration of DL algorithms for clinical applications have motivated the development of new methodological frameworks for *distributed collaborative learning*. These techniques allow collaborations to benefit from training DL models of their collective datasets without explicitly sharing the data (13). This is achieved by exchanging model weights or intermediate outputs during the training process. In this article, we describe popular methods for training distributed DL models in the context of health care institutions, showcase several real-world examples of their application in successful

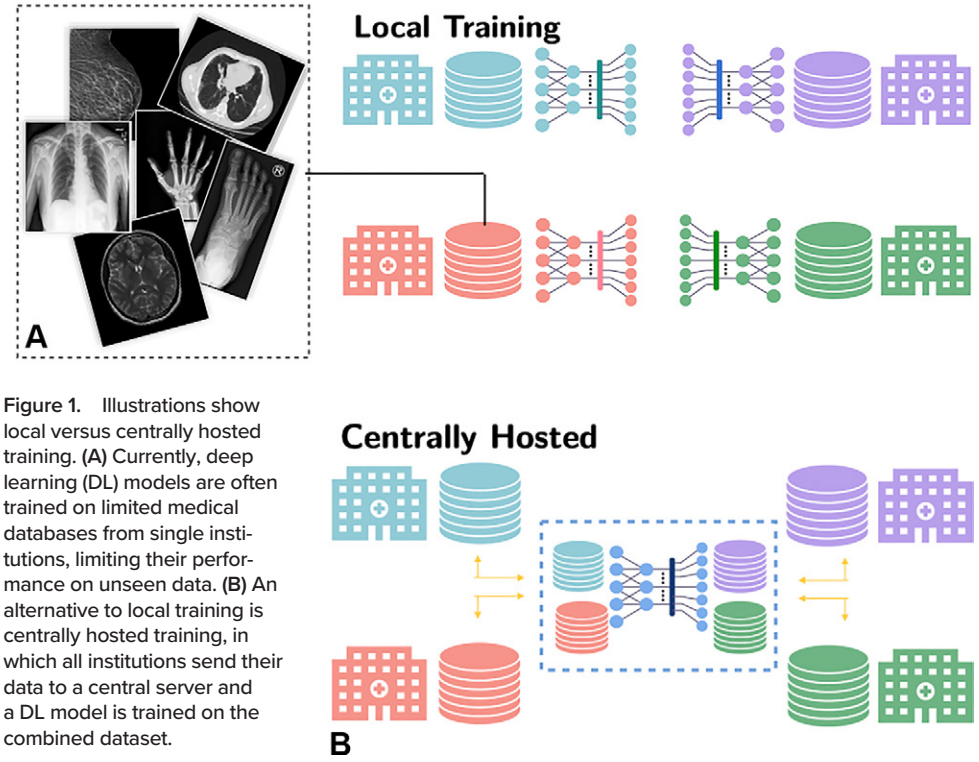

Figure 1. Illustrations show local versus centrally hosted training. (A) Currently, deep learning (DL) models are often trained on limited medical databases from single institutions, limiting their performance on unseen data. (B) An alternative to local training is centrally hosted training, in which all institutions send their data to a central server and a DL model is trained on the combined dataset.

collaborations, and highlight publicly available frameworks that can be used for collaborative learning. We also discuss some key challenges that remain in distributed DL collaborations and explore a few promising future research directions.

Methods for Distributed Collaborative Learning

Model Ensembling

There are several methods to perform distributed collaborative learning. The simplest approach is based on the concept of *model ensembling* (Fig 2A), in which multiple individual models are trained and then aggregated to improve accuracy over that of any single model alone (14). In a multi-institutional setting, each institution could train its own local version of the same model and then share the trained models with other institutions. To make a prediction with the ensemble of models, the input is fed into each of the models and then aggregated to determine the final prediction. This aggregation is commonly based on either majority voting or averaging. In the former method, the prediction from each model serves as its vote, and the final output prediction is the one receiving the most votes. However, in the latter method, predictions from each model are averaged to determine the final output.

Both of the ensembling approaches can be quite effective. Results of a 2019 study (15) showed that ensembling multiple decorrelated outputs from different models can be more accurate than any single model. One key advantage of ensembling is its simplicity; the only step that must be coordinated is the sharing of the different models, which can be performed asynchronously and without a central server. A potential disadvantage of the ensemble approach is the requirement that each institution has a sufficient amount of data to train a high-quality local model, because ensembling with an inadequately trained model results in ensemble predictions with low accuracy.

Federated Averaging

A popular approach for *federated learning* (16) is federated averaging, which involves training a local copy of the model at each institution that is then aggregated at a central server, usually by averaging the weights of all the local models to form a global model (Fig 2B). This global version of the aggregated model is then distributed back to each institution's server, where this process of training and aggregation is repeated for a number of iterations. A central server is typically needed for coordination among institutions to ensure that the averaging step is only performed after each institution has finished training its local model. One advantage of this technique is that each institution's model can be trained in parallel so that federated learning scales to a large number of institutions. In other settings, the algorithm has been applied to tens of thousands of clients (eg, mobile devices). However, participating institutions with long communication or training times can hold up the entire procedure. Methods to determine appropriate waiting times for each institution and to handle missing institutions from aggregation rounds have been developed (17).

Split Learning

Split learning (18) (Fig 2D) is another approach to distributed collaborations that also involves the use of a central server. Instead of training an entire copy of the model at each institution, only part of the model is trained locally (usually the initial and final layers of the deep neural network), while the rest of the model is trained on the central server (the middle layers of the deep neural network). For each institution, the

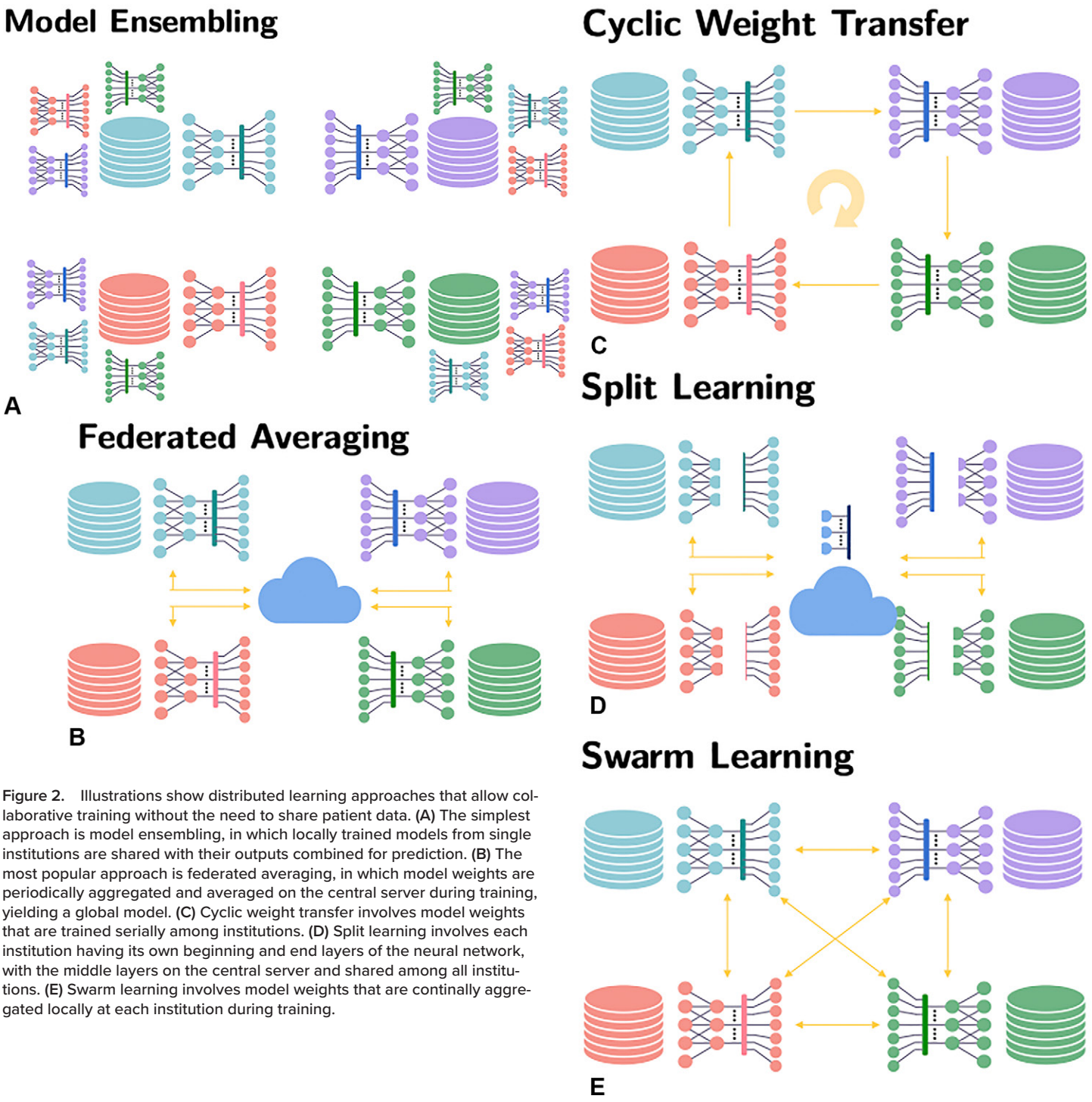

Figure 2. Illustrations show distributed learning approaches that allow collaborative training without the need to share patient data. **(A)** The simplest approach is model ensembling, in which locally trained models from single institutions are shared with their outputs combined for prediction. **(B)** The most popular approach is federated averaging, in which model weights are periodically aggregated and averaged on the central server during training, yielding a global model. **(C)** Cyclic weight transfer involves model weights that are trained serially among institutions. **(D)** Split learning involves each institution having its own beginning and end layers of the neural network, with the middle layers on the central server and shared among all institutions. **(E)** Swarm learning involves model weights that are continally aggregated locally at each institution during training.

respective input data are passed through the initial layers of the local model to generate an intermediate output that is shared with the central server to be further trained with a global model. This approach is often useful for vertically partitioned data, whereby institutions share the same patients but differ in the feature space (19). In medical imaging, such a collaboration may exist if a patient regularly visits one clinic but undergoes radiologic screening at another hospital and genomic sequencing at yet another clinic's laboratory.

One key advantage of this approach is that much of the intensive computations in the intermediate model layers can be offloaded to the central server to decrease hardware requirements for each local institution. In addition, split learning may be more efficient from a data communication standpoint when compared with federated learning when a large number of institutions are involved (20), because only a small portion of the entire model must be transmitted to and from the central server. Also, the split learning network can be trained in such a way that the intermediate output is decorrelated from its input, providing a privacy-preserving advantage (21). The disadvantage of split learning is an increase in training time, because only one institution can train the global model at a time, and global model updates are performed for every batch of data during training (rather than at the end of each round

of training as in federated learning. Efforts have been made to combine split learning with federated learning to avoid these training inefficiencies (22).

Weight Transfer

Although federated learning and split learning have many advantages, they require a central server to aggregate local model updates, which entails trusting that the central server has not been compromised by a malicious third party. To avoid the security concerns of maintaining a central server in multi-institutional collaborations, decentralized approaches have been developed. One such approach is *single weight transfer* (23), in which a model is trained at one institution for a fixed number of iterations before passing the trained model to the next institution (in a process similar to fine-tuning a model in transfer learning). This process is repeated until the model has been trained at all institutions. This is similar in concept to transfer learning and has similar weaknesses regarding the risk of overfitting and catastrophic forgetting. To help mitigate these weaknesses, the cycle of passing models between institutions can be repeated in a process called *cyclic weight transfer* (23) (Fig 2C). This modified approach has empirically demonstrated that frequent transferring of the model (ie, increased weight transfer frequency) to the next institution results in better performance of the final model. The downside of more frequent weight transfer is that there is a higher communication cost, which can be a bottleneck, because typically, federated learning systems are limited by an upload bandwidth of 1 MB per second or less (16). Unstable connections also limit the number of active clients in a round, further making communication costs a principal constraint.

Swarm Learning

Another decentralized approach is *swarm learning* (Fig 2E), in which use of a central server is avoided to organize communication and weight sharing by using a network of nodes, each representing separate institutions. This approach allows the dynamic addition of new participating institutions and automatic handling of synchronization of model weights and parameters between nodes.

Real-World Examples of Distributed DL for Health Care Applications

Although distributed learning is extensively used in many commercial use cases such as mobile devices, more regulated industries such as finance and health care have been slower to adopt these techniques for their machine learning applications (24). In addition to the challenges of medical data privacy, distributed learning for medical applications usually involves a consortium of health care institutions working collaboratively to develop DL tools for their shared benefit.

The complexity of coordination in a multi-institutional consortium is a challenge. For instance, it is difficult to come to an agreement on the approach to the clinical problem (ie, choosing how to preprocess medical datasets among institutions, including the interoperability of data or annotations). Harmonizing is another challenge if and when disagreement occurs between participating institutions regarding medical classification labels. For example, some institutions may have different criteria or sensitivity values for what they consider "abnormal." In addition, there can be substantial variability in interpretation among individual radiologists, even when they use standardized criteria (25). Issues regarding establishment of proper infrastructure to enable secure connections among sites also must be addressed to ensure that patient privacy is not compromised. Despite all of these challenges concerning health care datasets, numerous efforts have been made to develop real-world multi-institutional collaborative learning platforms.

Distributed learning has shown a high potential for various classification tasks (26). Roth et al (27) found that institutions from multiple countries jointly trained a DL model to assess breast density on mammograms using a distributed learning framework to provide communication between servers. Using a federated learning approach, they showed an increase in performance of 6.3% over that of models trained with only one institution's data and a 45.8% relative improvement in the models' generalizability over external test data. Authors of another study (28) leveraged federated learning to train a model using chest x-rays and electronic medical record data from 20 institutions to predict future oxygen requirements of symptomatic patients with COVID-19. They reported an increase in the generalizability of 38% over that of single-site models. Warnat-Herresthal et al (29) used swarm learning for classification tasks for four different medical applications: COVID-19, tuberculosis, leukemia, and lung abnormalities. The authors ensured privacy with a blockchain-based network protocol for communication. Using 16 400 blood transcriptomes from 127 clinical studies and 95 000 chest x-ray images, the study showed that swarm learning outperforms models that were developed independently at individual sites while satisfying local confidentiality regulations.

Distributed learning was also found to be successful for segmentation tasks. Sarma et al (30) developed a distributed learning model for whole prostate segmentation using training data from three different institutions. Not only did the distributed learning model result in superior performance compared with locally trained models, but it also improved performance in an external challenge dataset that was not used for training. In another seminal work (31), the authors used federated learning for training a brain tumor segmentation model using brain MRI datasets from 10 different institutions, achieving a dice score that was 99% of the dice score of a model trained on centrally hosted data. They also showcased how local models failed when tested on datasets from other institutions, while the distributed learning model, taking cues from an ample and diverse training dataset, generalized extremely well on almost all institutional datasets.

Open-Source Distributed Learning Frameworks

To facilitate and standardize federated learning for health care applications, open-source software platforms (32) are being actively developed (32–34), including the Federated Learning Application Runtime Environment (FLARE; Nvidia) (35), Open Federated Learning (Intel Labs) (36), and Flower (Adap) (37). The release of these frameworks has the potential to

greatly accelerate the adoption and use of institutional cooperation to develop robust and performant DL-based tools in the biomedical community. FLARE is an open-source software platform for distributed learning. FLARE currently supports workflows that allow federated averaging and cyclical weight transfer approaches as well as privacy-preserving features such as differential privacy and homomorphic encryption (ie, an encryption scheme that allows mathematical operations to be performed on data without the need for decryption). Extensive collaborations between industry and research partners have led to efforts such as Open Federated Learning, which is a project by Intel Corporation and the University of Pennsylvania, to support federated learning for segmentation of brain tumors. Flower is another federated learning framework developed by Adap that allows large-scale execution of heterogeneous clients (ie, up to 15 million edge devices on a pair of graphics processing unit hardware). While the previous platforms are built on top of commonly used DL frameworks (eg, usually TensorFlow [Google Brain Team] and PyTorch [Meta AI]), these are efforts to build distributed learning capabilities directly into these existing DL frameworks to facilitate further research of new techniques. TensorFlow Federated provides a low-level interface to build a custom communication protocol layer as well as a higher-level interface to easily adapt existing TensorFlow and Keras DL (François Chollet) models.

Although they are not as common as frameworks developed for federated learning, open-source platforms adaptable to multiple DL frameworks are underway for other distributed learning methods such as swarm learning (38). Hewlett Packard has developed a swarm learning framework that enables Tensorflow, Keras (*https://keras.io/guides/*), or Pytorch DL models to be deployed onto an Ethereum blockchain network. The Pysyft (*https://github.com/OpenMined/PySyft*) (33) library allows training on private data by using encrypted communication techniques (secure multiparty communication and homomorphic encryption) with federated learning and differential privacy. Fed-BioMed (Inria) (39) is an open-source project that provides a secure and deployable production-ready environment specifically tailored for the biomedical industry. These frameworks represent a sampling of the most popular open-source options used in medical imaging research.

Challenges

Distributional Shifts from Institutional Heterogeneity

To safely deploy clinical DL-based tools, it is important to recognize scenarios that can cause a model to have poor predictive accuracy (40). One common failure mode known to affect machine learning models is when there exists a mismatch between the data on which the model was developed and the data that the model actually encounters during deployment. This gap in data distributions between the development phase and the deployment phase may be broadly categorized into three main categories (41): *covariate shift*, *label shift*, and *concept shift*.

Covariate shift is defined as a change in P(X), where P(X) denotes the probability distribution of input images. Under covariate shift, the distribution of the input data may change, but the conditional probability of a target label, given a particular input, remains the same. A covariate shift often occurs when the cohort selected to train the model is not representative of the target population (similar to sampling bias). The shift in patient population demographics such as age and sex (Fig. 3A) or the differences in image acquisition such as resolution, hardware, modalities, sequences, or use of contrast material may also include covariate shift.

Label shift is defined as a change in P(Y), where P(Y) denotes the probability distribution of target labels or annotations. A label shift occurs when the probability of the target label changes. A label shift can manifest through differences in the disease prevalence, characteristics, or severity (eg, during the different phases and variants of the COVID-19 pandemic) (Fig 3B). Furthermore, annotations may vary among experts (radiologists) (Fig 3B), creating label shift.

Concept shift is defined as change in P(Y|X), where P(Y|X) denotes the conditional probability distribution of target labels (Y) given the input images (X). A concept shift occurs when the distribution of the input remains the same, but the conditional probability of the target label for any given input changes. Specifically, a concept shift is characterized by changes in the statistical properties of the target variables after the deployment of the model. A concept shift may occur if the efficacy rate of a new intervention affects certain groups of patients differently. A concept shift often occurs because of exogenous changes in the environment. For example, changing the diabetes diagnostic criteria for fasting plasma glucose may change the meaning of automatically derived labels. Hence, the task now requires retraining on the new data, which, in turn, alters the decision (classification or regression) function (Fig. 3C).

In light of the many modes of distribution shifts, curation and preprocessing of datasets are critical in mitigating challenges of data heterogeneity and preventing the loss of model performance. For example, before expert clinicians label the data, they must agree on standardized classification criteria to reduce interrater variability. In addition, data can be preprocessed using image processing techniques (eg, image intensity normalization) to harmonize the data among different imaging units and acquisition settings.

Distribution shift is a major challenge for distributed learning, because collaborative learning among institutions with different data distributions can negatively affect performance. When data distributions differ among institutions, this is referred to as non–independent and identically distributed (non-IID) data, whereas independent and identically distributed (IID) data are the ideal scenario. Authors of several studies (42,43) have shown that distributed learning performs worse in non-IID settings than it does in IID settings. Improving the performance of distributed collaborative learning in non-IID settings is an area of active research.

Data differences among institutions result in large gaps in the parameter space between the local optima of each institution and the global optima of the collaboratively trained model. Karimireddy et al (44) formalized this concept as *institutional drift*, whereby the local updates of each institution drift apart with respect to the server aggregated weights,

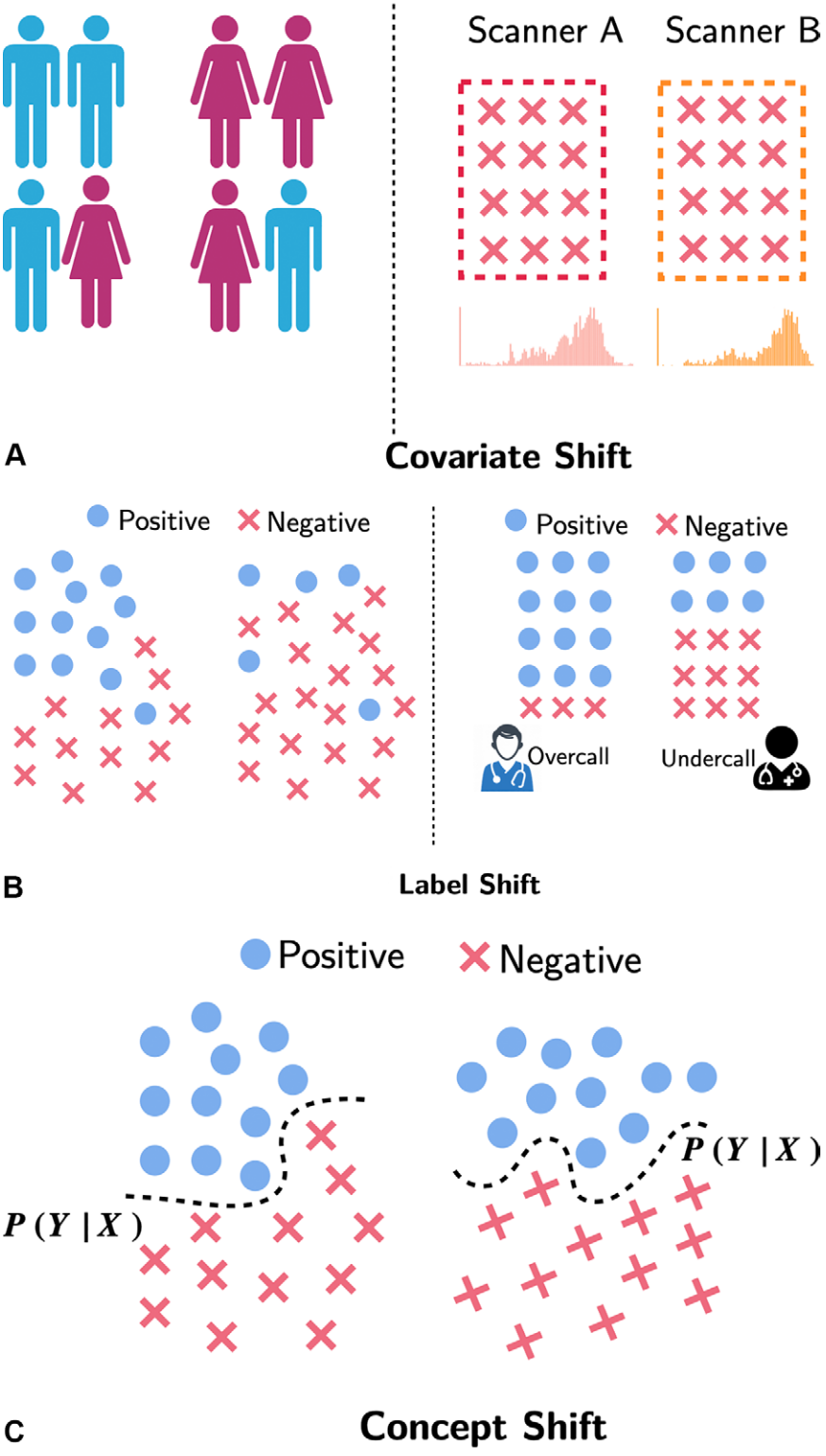

Figure 3. Illustration shows that data heterogeneity among institutions can make training collaborative learning algorithms challenging. This heterogeneity can include factors such as **(A)** covariate shift, **(B)** label shift, and **(C)** concept shift.

which results in slow, unstable convergence. Karimireddy et al propose a variance reduction technique to regularize institutional updates and remove unnecessary data biases among institutions. Acar et al (45) developed a method called FedDYN (federated learning based on dynamic regularization), where each participating institution dynamically updates a regularizer to align the local model with the global model. FedProx (46) imposes a squared penalty to the distance between the data in the server model and that in the model at each institution. In a complementary work, Wang et al (47) use reinforcement learning to select a subset of institutions to participate in any given round On the basis of the test accuracy achieved by the global model to counterbalance the biases introduced by training with non-IID data among institutions. Li et al (48)

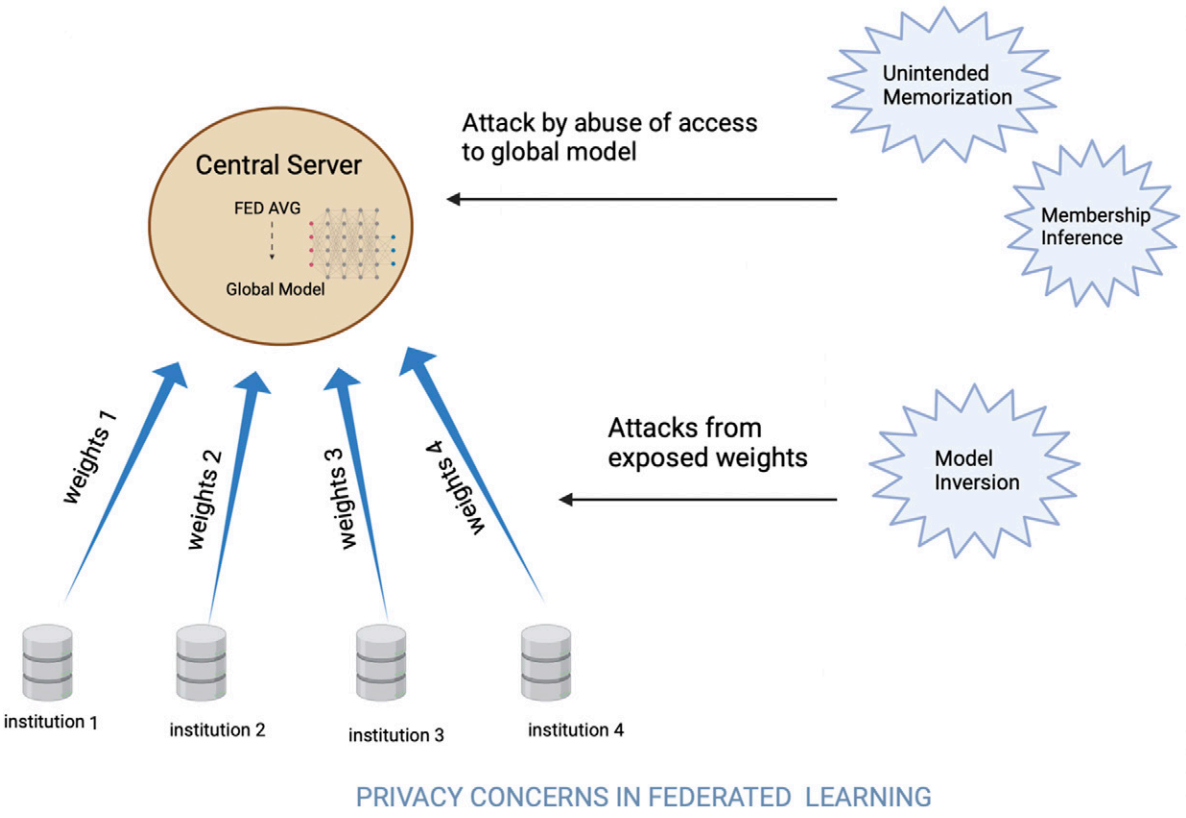

Figure 4. Illustration shows that although collaborative learning approaches preclude the need to share raw patient data, they do not necessarily guarantee the protection of patient privacy. Specifically, access to the global model's weights allows attacks such as unintended memorization and membership inference. Similarly, access to the model weight updates from institutions allows attacks such as model inversion. *FED AVG* = federated averaging.

proposed a bilevel optimization problem by incorporating a regularization term that constrains personalized models to be close to the global model.

Although most techniques have been focused on addressing the challenge of institutional heterogeneity through modifying a loss function or by solving a composite optimization problem (44,47,49–51), few authors (52,53), to our knowledge, have tried to model the underlying data-generating mechanism at each institution. The extra gradient methods we have described are not explicitly optimized to help determine causal features and thus may fail when introduced with out-of-distribution examples outside the aggregated distribution of the institutions (44,49–51,54). Specifically, when heterogeneity stems from selection bias, data acquisition, variations in patient population, and disease prevalence, studying the underlying causal structure provides tools to possibly mitigate them and train robust algorithms by using suitable strategies of annotation and machine learning frameworks.

Castro et al (55) discussed how causal considerations can have profound consequences on the soundness of the machine learning approach used and the resulting analysis, specifically in medical imaging. However, to our knowledge, few scientific works have been incorporated using principles of causal reasoning to improve distributed DL optimization in non-IID settings. Tenison et al (53) used the concept of gradient masking (56) adapted for causal learning. Gradient masking completely reduces those gradients in disagreement among institutions using a logical "and" operation, and updates only in the directions that are in agreement with those of a large fraction of institutional model gradients. In another work, Francis et al (52) adapted *invariant risk minimization* (57) to learn an invariant representation among institutions. Such a representation is agnostic to noncausal features among clients and elicits an invariant predictor that is simultaneously optimal for all institutions. Despite a surge of interest in this domain, the causal reasoning and invariant learning approaches lack theoretical guarantees of achieving the global minima and can face serious failure modes (58,59).

Privacy Concerns

Distributed DL approaches may still pose risks to data privacy even without explicitly sharing the data by exposing information contained in the model parameters, predictions, or gradients (60) (Fig 4). For example, *model inversion* attacks have been shown to reconstruct sensitive input data from the predicted output of a trained model. In much of the available literature (61), authors explore model inversion broadly under two possible attack scenarios: *(a)* when some prior knowledge of nonsensitive features of input data is known and *(b)* when no information about the input is known. The former case has been shown to allow considerably high accuracy of inversion attacks (62). Authors of another study (63) showed reasonable success in reidentification attack (ie, model inversion) without knowledge of input features for the specific case of facial reconstruction. Another type of privacy threat is *membership inference* (64,65). The goal of this attack is to determine to which institution a particular data point belongs. There are two classes of membership inference attacks: white-box and black-box attacks. In a white-box attack, the attacker is assumed to have full knowledge or access to the model, whereas in a black-box attack, access to the model is limited

to only observing the predicted outputs. Although simpler machine learning models are more susceptible to membership inference attacks under a black-box setting, success in the case of commonly used DL architectures has also been demonstrated in white-box attack scenarios (66).

Because DL algorithms can learn a larger set of highly complex functions, they also tend to learn sensitive features or correlations inadvertently from the input data, which may be exploited by an attacker to gain information about the original dataset. To address these vulnerabilities, several cryptographically based approaches have been developed to mitigate potential privacy issues. For example, in collaborative learning, *homomorphic encryption* (67) can be used to allow some types of mathematical operations to be performed on encrypted data, which can ultimately be used to share model parameters or gradients securely to protect against attacks that exploit unencrypted model information. However, homomorphic schemes for federated learning have their own limitations when applied in this way, specifically the increased computation time of encrypting and decrypting data and the increased size of encrypted messages (ie, model weights), which becomes difficult to handle with the increasing complexity of DL models. Another privacy technique similar to homomorphic encryption for distributed DL is to use *secure multiparty computation* (SMPC) (68), which establishes protocols that ensure secure computation of a common function of interest (eg, averaging in the case of federated averaging) without revealing the private details of data coming from different collaborating institutions. Other privacy preservation techniques include theoretical frameworks such as *differential privacy* (69), which provides statistical guarantees that ensure the privacy of any independent patient. This is accomplished by introducing noise by means of clipping the gradients in a way that it prevents any individual data point from leaking too much information about itself to the model that could subsequently be used to identify that patient.

In addition to cryptographically based approaches in the software, recent hardware advancements allow increased privacy, with specialized technology such as *confidential computing*. This approach involves processing information in secure and isolated blocks of hardware (ie, central processing unit– or graphics processing unit–based). Such secure blocks of the hardware are called the *trusted execution environment,* which enables the processing of data and information in an extremely secure and fast manner.

Fairness in Distributed Learning

Fairness can be broadly classified into two categories: performance fairness (Fig 5A) and collaboration fairness (Fig 5B). *Performance fairness* refers to disparities in model performance between different subgroups such as patient demographics or institutions. This difference can stem from measurement bias, underrepresentation of certain groups, or heterogeneity in data distribution between institutions (70,71). *Collaboration fairness* refers to the differential contribution of each institution toward the final model. Specifically, an institution that contributes larger amounts of and higher-quality data is expected to add more value to the collaborative learning consortium than other institutions do. Devising an appropriate reward structure on the basis of a relative attribution from institutions is crucial to providing incentives for distributed learning collaborations.

Regarding performance fairness, many existing studies (5,72) are focused only on uniform performance throughout institutions. However, from the perspective of health care disparities, mitigation rather than propagation of existing inequities is important. More studies are needed to achieve fair distributed learning performance throughout different subgroups such as sex, demographic location, socioeconomic status, age, and race. To date, only a few works (73–76) have addressed the challenge of subgroup fairness related to sensitive demographic attributes such as patient sex and race. Lu et al (73) applied quantification methods for distribution-free uncertainty such as conformal prediction to provide equal coverage guarantees for subgroups with different skin tones for classification of skin lesions. Conformal predictions have also been extended to distributed DL contexts for medical imaging applications (74) . Zhang et al (75) proposed a method that uses a reward mechanism to facilitate fairness among various groups using reinforcement learning with a reward function aimed to maximize the accuracy of the global model and minimize the discrimination index of demographic groups using a trade-off factor. Du et al (76) developed a fairness-aware distributed learning framework that achieves high accuracy and fairness with data from unknown testing institutions. They defined a weight for each data sample and incorporated this value in both the loss function and the fairness constraint. Although both of the previous methods required each institution to share the statistics of the sensitive attributes with the server, Ezzeldin et al (77) proposed the use of a flexible local debiasing strategy at each institution, yielding high fairness improvements under higher data heterogeneity. In their framework, the aggregated weights corresponding to each institution were updated adaptively in each round so that those institutions with a higher match between the global fairness measure (at the server) and the local fairness measure were given higher weighting.

Authors of several scientific works (70,78,79) also aimed to achieve collaborative fairness in distributed learning. The basis of achieving this fairness is an accurate valuation of the contributed data at each institution. One such approach is to use Shapley values, which stem from game theory. The idea of Shapley values (78) is to quantify the contribution of each player (ie, participating institution) in the context of a cooperative game. Once the value of data at each institution is quantified, each institution can be rewarded commensurately, such as financial rewards for participation or resulting commercial benefits from derived intellectual property. As an alternative, Lyu et al (70) proposed allowing participants to receive models with performance commensurate with their contributions. This is achieved by enforcing the participants to converge to different models, depending on their contribution toward federation. This contribution is measured in terms of the quality of the gradient updates

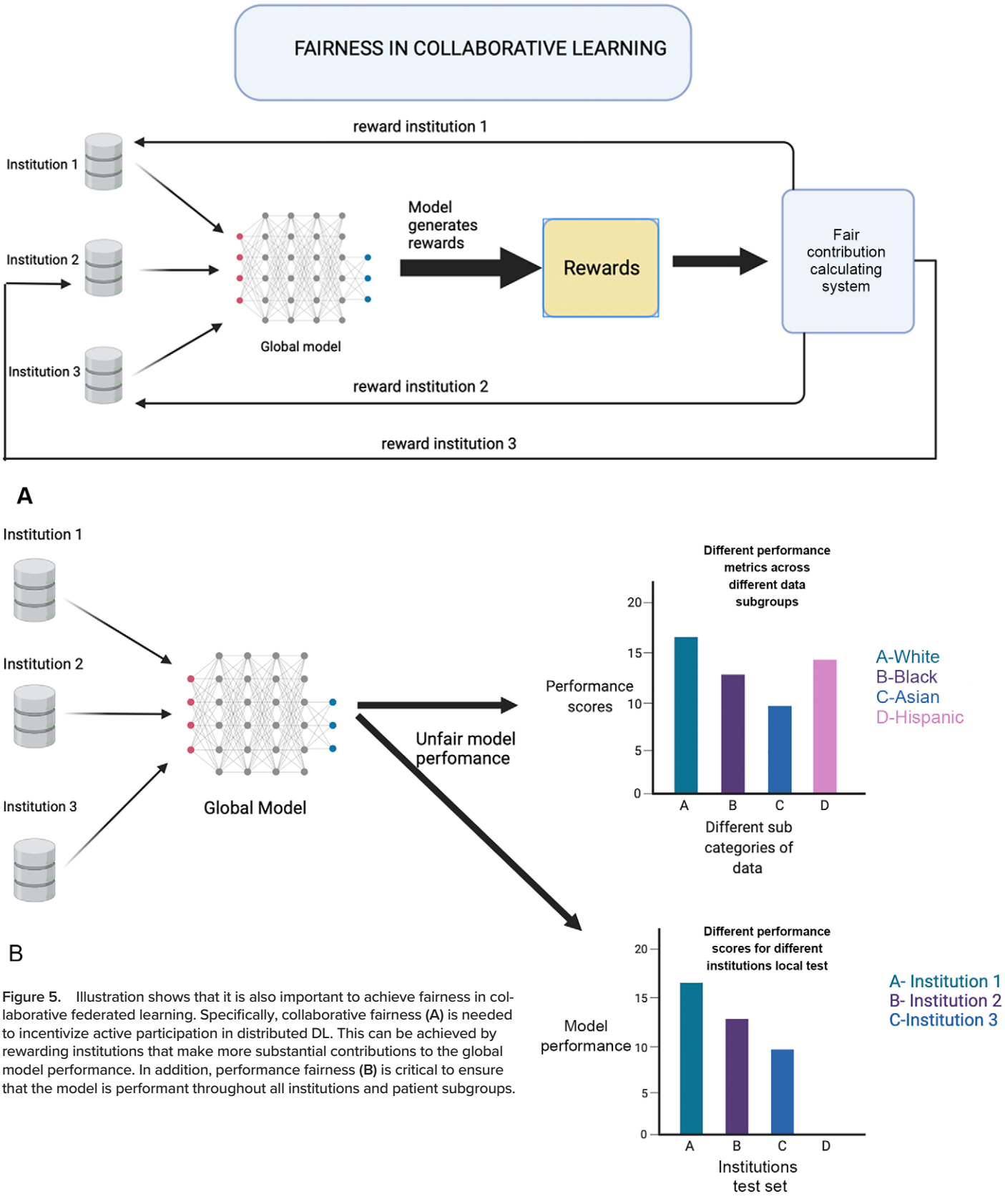

Figure 5. Illustration shows that it is also important to achieve fairness in collaborative federated learning. Specifically, collaborative fairness **(A)** is needed to incentivize active participation in distributed DL. This can be achieved by rewarding institutions that make more substantial contributions to the global model performance. In addition, performance fairness **(B)** is critical to ensure that the model is performant throughout all institutions and patient subgroups.

that an institution sends to the global model. Yu et al (79) modified the training optimization objective and incorporated waiting times and costs in the model in addition to the contributions of each institution.

Conclusion

Although DL algorithms have shown remarkable potential in automating various tasks for medical imaging and radiologic reports, models trained on limited data from single institutions

are often brittle and can underperform at external institutions. Distributed learning represents an exciting advance in multi-institutional collaborations to jointly build DL models without the need to share private patient data. Distributed learning has already shown promising results for various classification and segmentation tasks. Some key challenges of collaborative learning include data distribution shift, privacy leakage, and fairness. As research in distributed learning continues, new technical breakthroughs may shift the collaborative paradigm in both radiology and other medical disciplines. To effectively and safely use distributed learning methods, users from both technical and clinical backgrounds should be aware of the benefits as well as the risks.

Author affiliations.—From the Athinoula A. Martinos Center for Biomedical Imaging, Department of Radiology, Massachusetts General Hospital, Harvard Medical School, 13th Street, Building 149, Room 2301, Charlestown, MA 02129 (S.G., S.K., K.C., C.L., P.S., J.K.C.); and Indian Institute of Technology Delhi, New Delhi, India (S.G., S.K.). Presented as an education exhibit at the 2021 RSNA Annual Meeting. Received April 28, 2022; revision requested July 6 and received August 4; accepted August 9. **Address correspondence to** J.K.C. (email: *jkalpathy-cramer@mgh.harvard.edu*).

Funding.—J.K.C. supported by grants from the National Institutes of Health (U01CA154601, U24CA180927, U01CA242879, and U24CA180918) and the National Science Foundation (NSF1622542). Study supported by the Athinoula A. Martinos Center for Biomedical Imaging at the Massachusetts General Hospital and the Center for Functional Neuroimaging Technologies and a P41 biotechnology resource grant from the National Institute of Biomedical Imaging and Bioengineering, National Institutes of Health (P41EB015896).

Disclosures of conflicts of interest.—**J.K.C.** Institutional support from Bayer, GE, Nvidia, and Nuance Communications. All other authors, the editor, and the reviewers have disclosed no relevant relationships.

References

1. Chartrand G, Cheng PM, Vorontsov E, et al. Deep Learning: A Primer for Radiologists. RadioGraphics 2017;37(7):2113–2131.
2. Kaushal A, Altman R, Langlotz C. Geographic Distribution of US Cohorts Used to Train Deep Learning Algorithms. JAMA 2020;324(12):1212–1213.
3. Zech JR, Badgeley MA, Liu M, Costa AB, Titano JJ, Oermann EK. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. PLoS Med 2018;15(11):e1002683.
4. Chang K, Beers AL, Brink L, et al. Multi-Institutional Assessment and Crowdsourcing Evaluation of Deep Learning for Automated Classification of Breast Density. J Am Coll Radiol 2020;17(12):1653–1662.
5. Kim DW, Jang HY, Kim KW, Shin Y, Park SH. Design Characteristics of Studies Reporting the Performance of Artificial Intelligence Algorithms for Diagnostic Analysis of Medical Images: Results from Recently Published Papers. Korean J Radiol 2019;20(3):405–410.
6. Larson DB, Magnus DC, Lungren MP, Shah NH, Langlotz CP. Ethics of using and sharing clinical imaging data for artificial intelligence: A proposed framework. Radiology 2020;295(3):675–682.
7. National Lung Screening Trial Research Team; Aberle DR, Adams AM, et al. Reduced lung-cancer mortality with low-dose computed tomographic screening. N Engl J Med 2011;365(5):395–409.
8. Medical Imaging Data Resource Center. https://www.midrc.org/. Updated December 1, 2022. Accessed December 6, 2022.
9. Menze BH, Jakab A, Bauer S, et al. The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS). IEEE Trans Med Imaging 2015;34(10):1993–2024.
10. Petersen RC, Aisen PS, Beckett LA, et al. Alzheimer's Disease Neuroimaging Initiative (ADNI): clinical characterization. Neurology 2010;74(3):201–209.
11. Kirkpatrick J, Pascanu R, Rabinowitz N, et al. Overcoming catastrophic forgetting in neural networks. Proc Natl Acad Sci U S A 2017;114(13):3521–3526.
12. Gupta S, Singh P, Chang K, et al. Addressing catastrophic forgetting for medical domain expansion. arXiv preprint arXiv:2103.13511. https://arxiv.org/abs/2103.13511. Posted March 24, 2021. Accessed December 6, 2022.
13. Rieke N, Hancox J, Li W, et al. The future of digital health with federated learning. NPJ Digit Med 2020;3(1):119.
14. Lee S, Purushwalkam S, Cogswell M, Crandall D, Batra D. Why M Heads are Better than One: Training a Diverse Ensemble of Deep Networks. arXiv preprint arXiv:1511.06314. http://arxiv.org/abs/1511.06314. Posted November 19, 2015. Accessed December 6, 2022.
15. Pan I, Thodberg HH, Halabi SS, Kalpathy-Cramer J, Larson DB. Improving Automated Pediatric Bone Age Estimation Using Ensembles of Models from the 2017 RSNA Machine Learning Challenge. Radiol Artif Intell 2019;1(6):e190053.
16. McMahan HB, Moore E, Ramage D, Hampson S, Agüera y Arcas B. Communication-Efficient Learning of Deep Networks from Decentralized Data. arXiv preprint arXiv:1602.05629. http://arxiv.org/abs/1602.05629. February 17, 2016. Revised February 28, 2016. Accessed December 6, 2022.
17. Park Y, Han DJ, Kim DY, Seo J, Moon J. Few-Round Learning for Federated Learning. In: Ranzato M, Beygelzimer A, Dauphin Y, Liang PS, Vaughan JW, eds. Advances in Neural Information Processing Systems. Curran Associates, 2021; 28612–28622. https://proceedings.neurips.cc/paper/2021/file/f065d878ccfb4cc4f4265a4ff8bafa9a-Paper.pdf. 2021. Accessed December 6, 2022.
18. Vepakomma P, Gupta O, Swedish T, Raskar R. Split learning for health: Distributed deep learning without sharing raw patient data. arXiv preprint arXiv:1812.00564. https://arxiv.org/abs/1812.00564. December 3, 2018. Accessed December 6, 2022.
19. Yang Q, Liu Y, Chen T, Tong Y. Federated Machine Learning: Concept and Applications. arXiv preprint arXiv:1902.04885. https://arxiv.org/abs/1902.04885. February 13, 2019. December 6, 2022.
20. Singh A, Vepakomma P, Gupta O, Raskar R. Detailed comparison of communication efficiency of split learning and federated learning. arXiv preprint arXiv:1909.09145. https://arxiv.org/abs/1909.09145. September 18, 2019. Accessed December 6, 2022.
21. Vepakomma P, Gupta O, Dubey A, Raskar R. Reducing leakage in distributed deep learning for sensitive health data.
22. Thapa C, Chamikara MAP, Camtepe S, Sun L. SplitFed: When Federated Learning Meets Split Learning. 2020.
23. Chang K, Balachandar N, Lam C, et al. Distributed deep learning networks among institutions for medical imaging. J Am Med Inform Assoc 2018;25(8):945–954.
24. Gu R, Yang S, Wu F. Distributed Machine Learning on Mobile Devices: A Survey. arXiv preprint arXiv:1909.08329. https://arxiv.org/abs/1909.08329. Posted September 18, 2019. Accessed December 6, 2022.
25. Spayne MC, Gard CC, Skelly J, Miglioretti DL, Vacek PM, Geller BM. Reproducibility of BI-RADS breast density measures among community radiologists: a prospective cohort study. Breast J 2012;18(4):326–333.
26. Lu C, Hanif A, Singh P, et al. Federated learning for multicenter collaboration in ophthalmology: improving classification performance in retinopathy of prematurity. Ophthalmol Retina 2022;6(8):657–663.
27. Roth HR, Chang K, Singh P, et al. Federated Learning for Breast Density Classification: A Real-World Implementation. Lect Notes Comput Sci 2020;12444:181–191.
28. 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.
29. Warnat-Herresthal S, Schultze H, Shastry KL, et al. Swarm Learning for decentralized and confidential clinical machine learning. Nature 2021;594(7862):265–270.
30. Sarma KV, Harmon S, Sanford T, et al. Federated learning improves site performance in multicenter deep learning without data sharing. J Am Med Inform Assoc 2021;28(6):1259–1264.
31. Sheller MJ, Edwards B, Reina GA, et al. Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data. Sci Rep 2020;10(1):12598.
32. Top 7 Open-Source Frameworks for Federated Learning. https://www.apheris.com/blog-top7-open-source-frameworks-for-federated-learning. Accessed December 6, 2022.
33. OpenMinded. Pysyft. https://github.com/OpenMined/PySyft. Accessed December 6, 2022.
34. Tensorflow. TensorflowFederated. https://www.tensorflow.org/federated. Accessed December 6, 2022.
35. NVIDIA. NVFlare. https://github.com/NVIDIA/NVFlare. Accessed December 6, 2022.
36. Reina GA, Gruzdev A, Foley P, et al. OpenFL: An open-source framework for Federated Learning. arXiv preprint arXiv:2105.06413. https://arxiv.org/abs/2105.06413. May 13, 2021. Accessed December 6, 2022.
37. Beutel DJ, Topal T, Mathur A, et al. Flower: A Friendly Federated Learning Research Framework. https://www.flower.dev. 2020. Accessed December 6, 2022.
38. HewlettPackard. HewlettPackard/swarm-learning. https://github.com/HewlettPackard/swarm-learning. 2022. Accessed December 6, 2022.
39. Silva S, Altmann A, Gutman B, Lorenzi M. Fed-BioMed: A General Open-Source Frontend Framework for Federated Learning in Healthcare. Lect Notes Comput Sci 2020;12444:201–210.

40. Lu C, Chang K, Singh P, et al. Deploying clinical machine learning? Consider the following... arXiv preprint arXiv:2109.06919. https://arxiv.org/abs/2109.06919. September 14, 2021. Accessed December 6, 2022.
41. Huyen C. Data Distribution Shifts and Monitoring. https://huyenchip.com/2022/02/07/data-distribution-shifts-and-monitoring.html. February 7, 2022. Accessed December 6, 2022.
42. Mohri M, Sivek G, Suresh AT. Agnostic Federated Learning. February 1, 2019. Accessed December 6, 2022.
43. Sattler F, Wiedemann S, Muller KR, Samek W. Robust and Communication-Efficient Federated Learning From Non-i.i.d. Data. IEEE Trans Neural Netw Learn Syst 2020;31(9):3400–3413.
44. Karimireddy, S. P., Kale, S., Mohri, M., Reddi, S. J., Stich, S. U., & Suresh, A. T. (2020). SCAFFOLD: Stochastic Controlled Averaging for Federated Learning. In H. Daume, & A. Singh (Eds.), 37th International Conference on Machine Learning, ICML 2020 (pp. 5088-5099). (37th International Conference on Machine Learning, ICML 2020; Vol. PartF168147-7). International Machine Learning Society (IMLS).
45. Acar DAE, Zhao Y, Navarro RM, Mattina M, Whatmough PN, Saligrama V. Federated Learning Based on Dynamic Regularization. https://arxiv.org/abs/2111.04263. November 8, 2021. Accessed December 6, 2022.
46. Li T, Sahu AK, Zaheer M, Sanjabi M, Talwalkar A, Smith V. Federated Optimization in Heterogeneous Networks. https://arxiv.org/abs/1812.06127. Published December 14, 2018. Accessed December 6, 2022.
47. Wang H, Kaplan Z, Niu D, Li B. Optimizing Federated Learning on Non-IID Data with Reinforcement Learning. In: IEEE INFOCOM 2020 - IEEE Conference on Computer Communications, Toronto, ON, Canada, July 6–9, 2020. Piscataway, NJ: IEEE, 2020; 1698–1707.
48. Li T, Hu S, Beirami A, Smith V. Ditto: Fair and Robust Federated Learning Through Personalization. https://arxiv.org/abs/2012.04221. December 8, 2020. Accessed December 6, 2022.
49. Johnson R, Zhang T. Accelerating Stochastic Gradient Descent using Predictive Variance Reduction. In: Burges CJ, Bottou L, Welling M, Ghahramani Z, Weinberger KQ, eds. Advances in Neural Information Processing Systems. Curran Associates, 2013. https://proceedings.neurips.cc/paper/2013/file/ac1dd209cbcc5e5d1c6e28598e8cbbe8-Paper.pdf. Accessed December 6, 2022.
50. Li T, Sahu AK, Zaheer M, Sanjabi M, Talwalkar A, Smith V. FedDANE: A Federated Newton-Type Method. https://arxiv.org/abs/2001.01920 . January 7, 2020. Accessed December 6, 2022.
51. Konečný J, McMahan HB, Ramage D, Richtárik P. Federated Optimization: Distributed Machine Learning for On-Device Intelligence. https://arxiv.org/abs/1610.02527. October 8, 2016. Accessed December 6, 2022.
52. Francis S, Tenison I, Rish I. Towards Causal Federated Learning For Enhanced Robustness and Privacy. https://arxiv.org/abs/2104.06557. April 14, 2021. Accessed December 6, 2022.
53. Tenison I, Francis S, Rish I. Gradient Masked Federated Optimization. https://arxiv.org/abs/2104.10322. Published April 21, 2021. Accessed December 6, 2022.
54. Liang X, Shen S, Liu J, Pan Z, Chen E, Cheng Y. Variance Reduced Local SGD with Lower Communication Complexity. https://arxiv.org/abs/1912.12844. December 30, 2019. Accessed December 6, 2022.
55. Castro DC, Walker I, Glocker B. Causality matters in medical imaging. Nat Commun. 2020 Jul 22;11(1):3673. doi: 10.1038/s41467-020-17478-w. PMID: 32699250; PMCID: PMC7376027.
56. Parascandolo G, Neitz A, Orvieto A, Gresele L, Schölkopf B. Learning explanations that are hard to vary. https://arxiv.org/abs/2009.00329. September 1, 2020. Accessed December 6, 2022.
57. Arjovsky M, Bottou L, Gulrajani I, Lopez-Paz D. Invariant Risk Minimization. July 5, 2019. Accessed December 6, 2022.
58. Rosenfeld E, Ravikumar P, Risteski A. The Risks of Invariant Risk Minimization. arXiv preprint arXiv:2010.05761. https://arxiv.org/abs/2010.05761. October 12, 2020. Accessed December 6, 2022.
59. Kamath P, Tangella A, Sutherland DJ, Srebro N. Does Invariant Risk Minimization Capture Invariance? January 24, 2021. Accessed December 6, 2022.
60. Li Z, Wang L, Chen G, Shafiq M, Gu Z. A Survey of Image Gradient Inversion Against Federated Learning. TechRxiv. https://www.techrxiv.org/articles/preprint/A_Survey_of_Image_Gradient_Inversion_Against_Federated_Learning/18254723/1. January 14, 2022. Accessed December 6, 2022.
61. Hidano S, Murakami T, Katsumata S, Kiyomoto S, Hanaoka G. Model inversion attacks for prediction systems: Without knowledge of non-sensitive attributes. In: Proceedings - 2017 15th Annual Conference on Privacy, Security and Trust (PST) 2017; 115–124.
62. Fredrikson M, Lantz E, Jha S, Lin S, Page D, Ristenpart T. Privacy in Pharmacogenetics: An End-to-End Case Study of Personalized Warfarin Dosing. Proc USENIX Secur Symp. 2014; Aug:17-32.
63. Zhang Y, Jia R, Pei H, Wang W, Li B, Song D. The Secret Revealer: Generative Model-Inversion Attacks Against Deep Neural Networks. November 17, 2019. https://arxiv.org/abs/1911.07135. Accessed December 6, 2022.
64. Melis L, Song C, De Cristofaro E, Shmatikov V. Exploiting Unintended Feature Leakage in Collaborative Learning. In: 2019 IEEE Symposium on Security and Privacy (SP), San Francisco, CA, May 19–23, 2019. Piscataway, NJ: IEEE, 2019; 691–706. https://doi.org/10.1109/SP.2019.00029. Accessed December 2022.
65. Shokri R, Stronati M, Song C, Shmatikov V. Membership Inference Attacks against Machine Learning Models. arXiv preprint arXiv:1610.05820. https://arxiv.org/abs/1610.05820. October 18, 2016. Accessed December 6, 2022.
66. Nasr M, Shokri R, Houmansadr A. Comprehensive Privacy Analysis of Deep Learning: Passive and Active White-box Inference Attacks against Centralized and Federated Learning. https://arxiv.org/abs/1812.00910. December 3, 2016. Accessed December 6, 2022.
67. Acar A, Aksu H, Uluagac AS, Conti M. A Survey on Homomorphic Encryption Schemes: Theory and Implementation. ACM Computing Surveys Association for Computing Machinery. arXiv preprint arXiv:1704.03578. https://arxiv.org/abs/1704.03578. Posted April 12, 2017. Accessed December 6, 2022.
68. Veeningen M, Chatterjea S, Horváth AZ, et al. Enabling analytics on sensitive medical data with secure multi-party computation. Stud Health Technol Inform 2018;247:76–80.
69. Lyu M, Su D, Li N. Understanding the Sparse Vector Technique for Differential Privacy. arXiv preprint arXiv:1603.01699. https://arxiv.org/abs/1603.01699. March 5, 2016. Accessed December 6, 2022.
70. Lyu L, Xu X, Wang Q, Yu H. Collaborative Fairness in Federated Learning. arXiv preprint arXiv:2008.12161. https://arxiv.org/abs/2008.12161. August 27, 2020. Accessed December 6, 2022.
71. Li T, Sanjabi M, Ai F, Beirami A, Smith V. Fair Resource Allocation in Federated Learning. arXiv preprint arXiv:1905.10497. https://arxiv.org/abs/1905.10497. May 25, 2019. Accessed December 6, 2022.
72. Daneshjou R, Smith MP, Sun MD, Rotemberg V, Zou J. Lack of Transparency and Potential Bias in Artificial Intelligence Data Sets and Algorithms: A Scoping Review. JAMA Dermatol. 2021 November 1;157(11):1362–1369.
73. Lu C, Lemay A, Chang K, Hoebel K, Kalpathy-Cramer J. Fair Conformal Predictors for Applications in Medical Imaging. arXiv preprint arXiv:2109.04392. https://arxiv.org/abs/2109.04392. Posted September 9, 2021. Accessed December 6, 2022.
74. Lu C, Kalpathy-Cramer J. Distribution-Free Federated Learning with Conformal Predictions. https://arxiv.org/abs/2110.07661. October 14, 2022. Accessed January 6, 2023.
75. Zhang DY, Kou Z, Wang D. FairFL: A Fair Federated Learning Approach to Reducing Demographic Bias in Privacy-Sensitive Classification Models. In: 2020 IEEE International Conference on Big Data (Big Data), Atlanta, GA, December 10–13, 2020. Piscataway, NJ: IEEE, 2020; 1051–1060.
76. Du W, Xu D, Wu X, Tong H. Fairness-aware Agnostic Federated Learning. arXiv preprint arXiv:2010.05057. https://arxiv.org/abs/2010.05057. October 10, 2020. Accessed December 6, 2022.
77. Ezzeldin YH, Yan S, He C, Ferrara E, Avestimehr S. FairFed: Enabling Group Fairness in Federated Learning. arXiv preprint arXiv:2110.00857. https://arxiv.org/abs/2110.00857. Posted October 2, 2021. Accessed December 6, 2022.
78. Wang T, Rausch J, Zhang C, Jia R, Song D. A Principled Approach to Data Valuation for Federated Learning. arXiv preprint arXiv:2009.06192. https://arxiv.org/abs/2009.06192. Posted September 14, 2020. Accessed December 6, 2022.
79. Yu H, Liu Z, Liu Y, et al. A Fairness-aware Incentive Scheme for Federated Learning. In: AIES '20: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, February 2020. New York, NY: Association for Computing Machinery, 2020; 393–399.