Article published online: 2022-01-21

Privacy-Preserving Artificial Intelligence Techniques in Biomedicine

Reihaneh Torkzadehmahani[1] Reza Nasirigerdeh[1,2] David B. Blumenthal[3] Tim Kacprowski[4,5] Markus List[6] Julian Matschinske[7,8] Julian Spaeth[7,8] Nina Kerstin Wenke[7,8] Jan Baumbach[7,8,9]

[1] Institute for Artificial Intelligence in Medicine and Healthcare, Technical University of Munich, Munich, Germany
[2] Klinikum Rechts der Isar, Technical University of Munich, Munich, Germany
[3] Department of Artificial Intelligence in Biomedical Engineering (AIBE), Friedrich-Alexander University Erlangen-Nürnberg (FAU), Erlangen, Germany
[4] Division of Data Science in Biomedicine, Peter L. Reichertz Institute for Medical Informatics of TU Braunschweig and Medical School Hannover, Braunschweig, Germany
[5] Braunschweig Integrated Centre of Systems Biology (BRICS), TU Braunschweig, Braunschweig, Germany
[6] Chair of Experimental Bioinformatics, Technical University of Munich, Munich, Germany

Address for correspondence Reihaneh Torkzadehmahani, MSc, Artificial Intelligence in Medicine and Healthcare, Technical University of Munich, Walther-von-Dyck Str. 4, 85748 Garching, Germany (e-mail: reihaneh.torkzadehmahani@tum.de).

[7] E.U. Horizon2020 FeatureCloud Project Consortium
[8] Chair of Computational Systems Biology, University of Hamburg, Hamburg, Germany
[9] Institute of Mathematics and Computer Science, University of Southern Denmark, Odense, Denmark

Methods Inf Med 2022;61:e12–e27.

Abstract

Background Artificial intelligence (AI) has been successfully applied in numerous scientific domains. In biomedicine, AI has already shown tremendous potential, e.g., in the interpretation of next-generation sequencing data and in the design of clinical decision support systems.

Objectives However, training an AI model on sensitive data raises concerns about the privacy of individual participants. For example, summary statistics of a genome-wide association study can be used to determine the presence or absence of an individual in a given dataset. This considerable privacy risk has led to restrictions in accessing genomic and other biomedical data, which is detrimental for collaborative research and impedes scientific progress. Hence, there has been a substantial effort to develop AI methods that can learn from sensitive data while protecting individuals' privacy.

Method This paper provides a structured overview of recent advances in privacy-preserving AI techniques in biomedicine. It places the most important state-of-the-art approaches within a unified taxonomy and discusses their strengths, limitations, and open problems.

Keywords
- privacy-preserving AI techniques
- federated learning
- biomedicine

Note: This work was done during the time Reihaneh Torkzadehmahani was a member of the FeatureCloud consortium and affiliated with the Chair of Experimental Bioinformatics, Technical University of Munich.

received
March 22, 2021
accepted after revision
September 18, 2021
published online
January 21, 2022

DOI https://doi.org/10.1055/s-0041-1740630.
ISSN 0026-1270.

Georg Thieme Verlag KG, Rüdigerstraße 14, 70469 Stuttgart, Germany

Conclusion As the most promising direction, we suggest combining federated machine learning as a more scalable approach with other additional privacy-preserving techniques. This would allow to merge the advantages to provide privacy guarantees in a distributed way for biomedical applications. Nonetheless, more research is necessary as hybrid approaches pose new challenges such as additional network or computation overhead.

Introduction

Artificial intelligence (AI) strives to emulate the human mind and to solve complex tasks by learning from available data. For many complex tasks, AI already surpasses humans in terms of accuracy, speed, and cost. Recently, the rapid adoption of AI and its subfields, specifically machine learning and deep learning, has led to substantial progress in applications such as autonomous driving,[1] text translation,[2] and voice assistance.[3] At the same time, AI is becoming essential in biomedicine, where big data in health care necessitates techniques that help scientists to gain understanding from it.[4]

Success stories such as acquiring the compressed representation of drug-like molecules,[5] modeling the hierarchical structure and function of a cell[6] and translating magnetic resonance images to computed tomography[7] using deep learning models illustrate the remarkable performance of these AI approaches. AI has not only achieved remarkable success in analyzing genomic and biomedical data,[8–18] but has also surpassed humans in applications such as sepsis prediction,[19] malignancy detection in mammography,[20] and mitosis detection in breast cancer.[21]

Despite these AI-fueled advancements, important privacy concerns have been raised regarding the individuals who contribute to the datasets. While taking care of the confidentiality of sensitive biological data is crucial,[22] several studies showed that AI techniques often do not maintain data privacy.[23–26] For example, attacks known as membership inference can be used to infer an individual's membership by querying over the dataset[27] or the trained model,[23] or by having access to certain statistics about the dataset.[28–30] Homer et al[28] showed that under some assumptions, an adversary (an attacker who attempts to invade data privacy) can use the statistics published as the result of genome-wide association studies (GWAS) to find out if an individual was a part of the study. Another example of this kind of attack was demonstrated by attacks on Genomics Beacons,[27,31] in which an adversary could determine the presence of an individual in the dataset by simply querying the presence of a particular allele. Moreover, the attacker could identify the relatives of those individuals and obtain sensitive disease information.[27,32] Besides targeting the training dataset, an adversary may attack a fully trained AI model to extract individual-level membership by training an adversarial inference model that learns the behavior of the target model.[23]

As a result of the aforementioned studies, health research centers such as the National Institutes of Health (NIH) as well as hospitals have restricted access to the pseudonymized data.[22,33,34] Furthermore, data privacy laws such as those enforced by the Health Insurance Portability and Accountability Act (HIPAA), and the Family Educational Rights and Privacy Act (FERPA) in the U.S. as well as the EU General Data Protection Regulation (GDPR) restrict the use of sensitive data.[35,36] Consequently, getting access to these datasets requires a lengthy approval process, which significantly impedes collaborative research. Therefore, both industry and academia urgently need to apply privacy-preserving techniques to respect individual privacy and comply with these laws.

This paper provides a systematic overview over various recently proposed privacy-preserving AI techniques in biomedicine, which facilitate the collaboration between health research institutes. Several efforts exist to tackle the privacy concerns in several domains, some of which have been examined in a couple of surveys.[37–39] Aziz et al[37] investigated previous studies which employed differential privacy and cryptographic techniques for human genomic data. Kaissis et al[39] briefly reviewed federated learning, differential privacy and cryptographic techniques applied in medical imaging. Xu et al[38] surveyed general solutions to challenges in federated learning including communication efficiency, optimization, as well as privacy and discussed possible applications including a few examples in health care. Compared with Aziz et al and Kaissis et al,[37,39] this paper covers a broader set of privacy preserving techniques including federated learning and hybrid approaches. In contrast with Xu et al[38] we additionally discuss cryptographic techniques and differential privacy approaches and their applications in biomedicine. Moreover, this survey covers a wider range of studies that employed different privacy-preserving techniques in genomics and biomedicine and compares the approaches using different criteria such as privacy, accuracy, and efficiency. It is notable that there are some hardware-based privacy-preserving approaches such as Intel Software Guard Extensions[40–42] and AMD memory encryption,[43] which allow for secure computation via secure hardware which are beyond the scope of this study.

The presented approaches are divided into four categories: cryptographic techniques, differential privacy, federated learning, and hybrid approaches. First, we describe how cryptographic techniques—in particular, homomorphic encryption (HE) and secure multiparty computation (SMPC)—ensure secrecy of sensitive data by carrying out computations on encrypted biological data. Next, we illustrate the differential privacy approach and its capability in quantifying individuals' privacy in published summary statistics of, for instance, GWAS data and deep learning models trained on

clinical data. Then, we elaborate on federated learning, which allows health institutes to train AIs locally and to share only selected parameters without sensitive data with a coordinator, who aggregates them and builds a global model. Following that, we discuss hybrid approaches which enhance data privacy by combining federated learning with other privacy-preserving techniques. We elaborate on the strengths and drawbacks of each approach as well as its applications in biomedicine. More importantly, we provide a comparison among the approaches with respect to different criteria such as computational and communication efficiency, accuracy, and privacy. Finally, we discuss the most realistic approaches from a practical viewpoint and provide a list of open problems and challenges that remain for the adoption of these techniques in real-world biomedical applications.

Our review of privacy-preserving AI techniques in biomedicine yields the following main insights: First, cryptographic techniques such as HE and SMPC, which follow the paradigm of "bring data to computation", are not computationally efficient and do not scale well to large biomedical datasets. Second, federated learning follows the paradigm of "bring computation to data" is a more scalable approach. However, its network communication efficiency is still an open problem and it does not provide privacy guarantees. Third, hybrid approaches that combine cryptographic techniques or differential privacy with federated learning are the most promising privacy-preserving AI techniques for biomedical applications, because they promise to combine the scalability of federated learning with the privacy guarantees of cryptographic techniques or differential privacy.

Cryptographic Techniques

In biomedicine and GWAS in particular, cryptographic techniques have been used to collaboratively compute result statistics while preserving data privacy.[40,44–56] These cryptographic approaches are based on HE[57–59] or SMPC.[60] There are different HE-based techniques such as partially HE (PHE)[58] and fully HE (FHE).[57] PHE allows either addition or multiplication operations to be performed on the encrypted data while using FHE both addition and multiplication operations can be applied. All HE-based approaches share three steps (►Fig. 1A):

1. Participants (e.g., hospitals or medical centers) encrypt their private data and send the encrypted data to a computing party.
2. The computing party calculates the statistics over the encrypted data and shares the statistics (which are encrypted) with the participants.
3. The participants access the results by decrypting them.

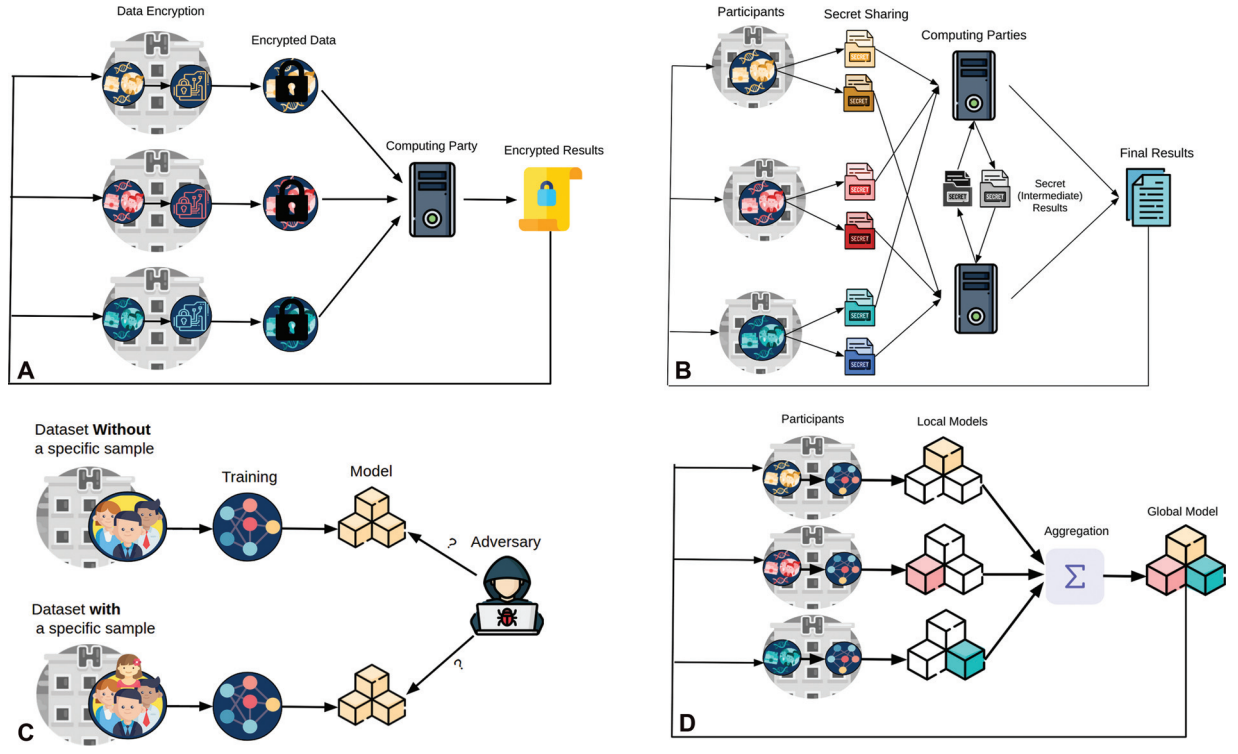

Fig. 1 Different privacy-preserving AI techniques: (**A**) *homomorphic encryption*, where the participants encrypt the private data and share it with a computing party, which computes the aggregated results over the encrypted data from the participants; (**B**) *secure multiparty computation* in which each participant shares a separate, different secret with each computing party; the computing parties calculate the intermediate results, secretly share them with each other, and aggregate all intermediate results to obtain the final results; (**C**) *differential privacy*, which ensures the models trained on datasets including and excluding a specific individual look statistically indistinguishable to the adversary; (**D**) *federated learning*, where each participant downloads the global model from the server, computes the local model given its private data and the global model, and finally sends its local model to the server for aggregation and for updating the global model. (**A**). Homomorphic encryption. (**B**). Secure multiparty computation. (**C**). Differential privacy. (**D**). Federated learning.

In SMPC, there are multiple participants as well as a couple of computing parties which perform computations on secret shares from the participants. Given M participants and N computing parties, SMPC-based approaches follow three steps (►Fig. 1B):

1. Each participant sends a separate and different secret to each of the N computing parties.
2. Each computing party computes the intermediate results on the M secret shares from the participants and shares the intermediate results with the other $N-1$ computing parties.
3. Each computing party aggregates the intermediate results from all computing parties including itself to calculate the final (global) results. In the end, the final results computed by all computing parties are the same and can be shared by the participants.

To clarify the concepts of secret sharing[61] and multiparty computation, consider a scenario with two participants P_1 and P_2 and two computing parties C_1 and C_2.[46] P_1 and P_2 possess the private data X and Y, respectively. The aim is to compute $X+Y$, where neither P_1 nor P_2 reveals its data to the computing parties. To this end, P_1 and P_2 generate random numbers R_X and R_Y, respectively; P_1 reveals R_X to C_1 and $(X-R_X)$ to C_2; likewise, P_2 shares R_Y with C_1 and $Y-R_Y$ with C_2; $R_X, R_Y, (X-R_X)$ and $(Y-R_Y)$ are secret shares. C_1 computes (R_X+R_Y) and sends it to C_2 and C_2 calculates $(X-R_X)+(Y-R_Y)$ and reveals it to C_1. Both C_1 and C_2 add the result they computed to the result each obtained from the other computing party. The sum is in fact $(X+Y)$, which can be shared with P_1 and P_2.

Notice that to preserve the data privacy, the computing parties C_1 and C_2 must be non-colluding. That is, C_1 must not send R_X and R_Y to C_2 and C_2 must not share $(X-R_X)$ and $(Y-R_Y)$ with C_1. Otherwise, the computing parties can compute X and Y, revealing the participants' data. In general, in an SMPC with N computing parties, data privacy is protected as long as most $N-1$ computing parties collude with each other. The larger the N, the stronger the privacy but higher the communication overhead and processing time. Another point is that, in addition to secret sharing, there are other transfer protocols in SMPC such as oblivious transfer[62] and garbled circuit[63] which is a two-party computation protocol in which each of the parties hold its private input and they jointly learn the output function describing the relation between their private inputs. Moreover, threshold cryptography combines a secret sharing scheme with cryptography to secretly share a key across distributed parties such that multiple parties (more than a threshold) must coordinate to encrypt/decrypt a message.[59,64] That is, threshold cryptography can be considered as the combination of the HE and SMPC methods.

Most studies use HE or SMPC to develop secure, privacy-aware algorithms applicable to GWAS data. Kim and Lauter[47] and Lu et al[49] implemented a secure^2 test and Lauter et al[48] developed privacy-preserving versions of common statistical tests in GWAS, such as the Pearson goodness of fit test, tests for linkage disequilibrium, and the Cochran Armitage trend test using HE. Kim et al[65] and Morshed et al[66] presented HE-based secure logistic and linear regression algorithms for medical data, respectively. Zhang et al,[53] Constable et al,[52] and Kamm et al[51] developed a SMPC-based secure χ^2 test. Shi et al[67] implemented a privacy-preserving logistic regression and Bloom[68] proposed a secure linear regression based on SMPC for GWAS data. Cho et al[44] introduced a SMPC-based framework to facilitate quality control and population stratification correction for large-scale GWAS and argued that their framework is scalable to one million individuals and half million single nucleotide polymorphisms (SNPs).

There are also other types of encryption techniques such as somewhat homomorphic encryption (SWHE),[57] which are employed to address privacy issues in genomic applications such as outsourcing genomic data computation to the cloud, and are not the main focus of this review. The main drawback of SWHE is that the number of successive addition and multiplication operations it can perform on the data are limited.[47] For more details, we refer to the comprehensive review by Mittos et al.[69]

Despite the promises of HE/SMPC-based privacy-preserving algorithms (►Table 1), the road for the wide adoption of HE/SMPC-based algorithms in genomics and biomedicine is long.[70] The major limitations of HE are few supported operations and computational overhead.[71] HE supports only addition and multiplication operations, and as a result, developing complex AI models with non-linear operations such as deep neural networks (DNNs) using HE is very challenging. Moreover, HE incurs remarkable computational overhead since it performs operations on encrypted data. Although SMPC is more efficient than HE from a computational perspective, it still suffers from high computational overhead,[72] which comes from processing secret shares from a large number of participants or large amount of data by a few computing parties.

Differential Privacy

One of the state-of-the-art concepts for eliminating and quantifying the chance of information leakage is *differential privacy*.[73–75] Differential privacy is a mathematical model that encapsulates the idea of injecting enough randomness or noise to sensitive data to camouflage the contribution of each single individual. This is achieved by inserting uncertainty into the learning process so that even a strong adversary with arbitrary auxiliary information about the data will still be uncertain in identifying any of the individuals in the dataset. This has become standard in data protection and has been effectively deployed by Google[76] and Apple[77] as well as agencies such as the United States Census Bureau. Furthermore, it has drawn the attention of researchers in privacy-sensitive fields such as biomedicine and health care[78–93].

Differential privacy ensures that the model we train does not overfit the sensitive data of a particular user. The model trained on a dataset containing information of a specific individual should be statistically indistinguishable from a model trained without the individual (►Fig. 1C). As an

Table 1 Literature for cryptographic techniques and differential privacy in biomedicine

Authors	Year	Technique	Model	Application
Kim and Lauter[47]	2015	HE	χ^2 statistics Minor allele frequency Hamming Distance Edit distance	Genetic associations DNA comparison
Lu et al[49]	2015	HE	χ^2 statistics D' measure	Genetic associations
Lauter et al[48]	2014	HE	D' and r^2 measure Pearson goodness-of-fit expectation maximization Cochran-Armitage	Genetic associations
Kim et al[65]	2018	HE	Logistic regression	Medical decision-making
Morshed et al[66]	2018	HE	Linear regression	Medical decision-making
Kamm et al[51]	2013	SMPC	χ^2 statistics	Genetic associations
Constable et al[52] Zhang et al[53]	2015 2015	SMPC	χ^2 statistics Minor allele frequency	Genetic associations
Shi et al[67]	2016	SMPC	Logistic regression	Genetic associations
Bloom[68]	2019	SMPC	Linear regression	Genetic associations
Cho et al[44]	2018	SMPC	Quality Control Population stratification	Genetic associations
Johnson and Shmatikov[78]	2013	DP	Distance-score mechanism p-value and χ^2 statistics	Querying genomic databases
Cho et al[95]	2020	DP	α-geometric mechanism	Querying biomedical databases
Aziz et al[79]	2017	DP	Eliminating random positions Biased random response	Querying genomic databases
Han et al[80] Yu et al[81]	2019 2014	DP	Logistic regression	Genetic associations
Honkela et al[82]	2018	DP	Bayesian linear regression	Drug sensitivity prediction
Simmons et al[83]	2016	DP	EIGENSTRAT Linear mixed model	Genetic associations
Simmons and Berger[84]	2016	DP	Nearest neighbor optimization	Genetic associations
Fienberg et al[85] Uhlerop et al[86] Yu and Ji[87] Wang et al[88]	2011 2013 2014 2014	DP	Statistics such as p-value, χ^2 and contingency table	Genetic associations
Abay et al[97]	2018	DP	Deep autoencoder	Generating artificial biomedical data
Beaulieu et al[98]	2019	DP	GAN	Simulating SPRINT trial
Jordon et al[99]	2018	DP	GAN	Generating artificial biomedical data

Abbreviations: DP, differential privacy; HE, homomorphic encryption, SMPC, secure multiparty computation.

example, assume that a patient would like to give consent to his/her doctor to include his/her personal health record in a biomedical dataset to study the coordination between age and cardiovascular disease. Differential privacy provides a mathematical guarantee which captures the privacy risk associated with the patient's participation in the study and explains to what extent the analyst or the potential adversary can learn about that particular individual in the dataset. Note that, differential privacy is typically employed for centralized datasets, where the output of the algorithm is perturbed with noise. However, SMPC and HE are leveraged for use cases where data are distributed across multiple clients, and carry out computation over the encrypted data or secret shares from the data of the clients. Formally, a randomized algorithm (an algorithm that has randomness in its logic and whose output can vary even on a fixed input) $A: D^n \rightarrow Y$ is (ε, δ)-differentially private if for all subsets $y \subseteq Y$ and for all adjacent datasets $D, D' \in D^n$ that differ in at most one record, then the following inequality holds:

$$Pr[A(D) \in y] \leq e^{\varepsilon} Pr[A(D') \in y] + \delta$$

Here, ε and δ are privacy loss parameters where lower values imply stronger privacy guarantees. δ is an exceedingly

small value (e.g., 10^{-5}) indicating the probability of an uncontrolled breach, where the algorithm produces a specific output only in the presence of a specific individual and not otherwise. ε represents the worst case privacy breach in the absence of any such rare breach. If you assume $\delta = 0$, you will have a pure (ε)-differentially private algorithm, while if you consider $\delta > 0$ to approximate the case in which pure differential privacy is broken, you will have an approximate (ε, δ)-differentially private algorithm.

Two important properties of differential privacy are composability[94] and resilience to post-processing. Composability means that combining multiple differentially private algorithms yields another differentially private algorithm. More precisely, if you combine k (ε, δ)-differentially private algorithms, the composed algorithm is at least ($k\varepsilon$, $k\delta$)-differentially private. Differential privacy also assures the resistance to post-processing theorem which states passing the output of a (ε, δ)-differentially private algorithm to any arbitrary randomized algorithm will still uphold the (ε, δ)-differential privacy guarantee.

The community efforts to ensure the privacy of sensitive genomic and biomedical data using differential privacy can be grouped into four categories according to the problem they address (►Table 1):

1. Approaches to querying biomedical and genomics databases.[78,79,93,95]
2. Statistical and AI modeling techniques in genomics and biomedicine.[80–84,92,96]
3. Data release, i.e., releasing summary statistics of a GWAS such as p-values and χ^2 contingency tables.[85–88]
4. Training privacy-preserving generative models.[97–99]

Studies in the first category proposed solutions to reduce the privacy risks of genomics databases such as GWAS databases and genomics beacon service.[100] The Beacon Network[31] is an online web service developed by the Global Alliance for Genomics and Health (GA4GH) through which the users can query the data provided by owners or research institutes, ask about the presence of a genetic variant in the database, and get a YES/NO as response. Studies have shown that an attacker can detect membership in the Beacon or GWAS by querying these databases multiple times and asking different questions.[27,101–103] Very recently, Cho et al[95] proposed a theoretical differential privacy mechanism to maximize the utility of count query in biomedical systems while guaranteeing data privacy. Johnson and Shmatikov[78] developed a differentially private query-answering framework. With this framework an analyst can retrieve statistical properties such as the correlation between SNPs and get an almost accurate answer while the GWAS dataset is protected against privacy risks. In another study, Aziz et al[79] proposed two algorithms to make the Beacon's response inaccurate by controlling a bias variable. These algorithms decide when to answer the query correctly/incorrectly according to specific conditions in the bias variable so that it gets harder for the attacker to succeed.

Some of the efforts in the second category addressed the privacy concerns in GWAS by introducing differentially private logistic regression to identify associations between SNPs and diseases[80] or associations among multiple SNPs.[81] Honkela et al[82] improve drug sensitivity prediction by effectively employing differential privacy for Bayesian linear regression. Moreover, Simmons et al[83] presented a differentially private EIGENSTRAT (PrivSTRAT)[104] and linear mixed model (PrivLMM)[105] to correct for population stratification. In another paper, Simmons et al[84] tackled the problem of finding significant SNPs by modeling it as an optimization problem. Solving this problem provides a differentially private estimate of the neighbor distance for all SNPs so that high scoring SNPs can be found.

The third category focused on releasing summary statistics such as p-values, χ^2 contingency tables, and minor allele frequencies in a differentially private fashion. The common approach in these studies is to add Laplacian noise to the true value of the statistics, so that sharing the perturbed statistics preserves privacy of the individuals. They vary in the sensitivity of the algorithm (that is, the maximum change on the output of an algorithm in presence or absence of a specific data point) and hence require different injected noise.[85,86,88]

The fourth category proposed novel privacy-protecting methods to generate synthetic health care data leveraging differentially private generative models (►Fig. 2). Deep generative models, such as generative adversarial networks (GANs),[106] can be trained on sensitive genomics and biomedical data to capture its properties and generate artificial data with similar characteristics as the original data.

Abay et al[97] presented a differentially private deep generative model, DP-SYN, a generative autoencoder that splits the input data into multiple partitions, then learns and simulates the representation of each partition while maintaining the privacy of input data. They assessed the performance of DP-SYN on sensitive datasets of breast cancer and diabetes. Beaulieu et al[98] trained an auxiliary classifier GAN (AC-GAN) in a differentially private manner to simulate the participants of the SPRINT trial (Systolic Blood Pressure Trial), so that the clinical data can be shared while respecting participants' privacy. In another approach, Jordon et al[99] introduced a differentially private GAN, PATE-GAN, and evaluated the quality of synthetic data on Meta-Analysis Global Group in Chronic Heart Failure (MAGGIC) and the United Network for Organ Transplantation (UNOS) datasets. Despite the aforementioned achievements in adopting differential privacy in the field, several challenges remain to be addressed. Although differential privacy involves less network communication, memory usage, and time complexity compared with cryptographic techniques, it still struggles with giving highly accurate results within a reasonable privacy budget, namely, intended ε and δ, on large-scale datasets such as genomics datasets.[37,107] In more detail, since genomic datasets are huge, the sensitivity of the applied algorithms on these datasets is large. Hence, the amount of distortion required for anonymization increases significantly, sometimes to the extent that the results will not be meaningful anymore.[108] Therefore, to make differential privacy more practical in the field, balancing a tradeoff

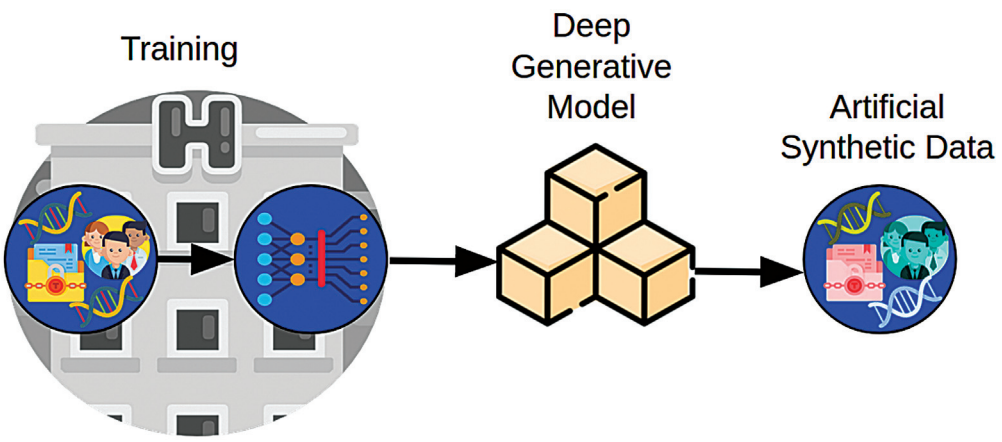

Fig. 2 Differentially private deep generative models: The sensitive data holder (e.g., health institutes) train a differentially private generative model locally and share just the trained data generator with the outside world (e.g., researchers). The shared data generator can then be used to produce artificial data with the same characteristics as the sensitive data.

between privacy and utility demands more attention than it has received[88,90–92].

Federated Learning

Federated learning[109] is a type of distributed learning where multiple clients (e.g., hospitals) collaboratively learn a model under the coordination of a central server while preserving the privacy of their data. Instead of sharing its private data with the server or the other clients, each client extracts knowledge (that is, model parameters) from its data and transfers it to the server for aggregation (►**Fig. 1D**).

Federated learning is an iterative process in which each iteration consists of the following steps[110]:

1. The server chooses a set of clients to participate in the current iteration of the model.
2. The selected clients obtain the current model from the server.
3. Each selected client computes the local parameters using the current model and its private data (e.g., runs gradient descent algorithm initialized by the current model on its local data to obtain the local gradient updates).
4. The server collects the local parameters from the selected clients and aggregates them to update the current model.

The data of the clients can be considered as a table, where rows represent samples (e.g., individuals) and columns represent features or labels (e.g., age, case vs. control). We refer to the set of samples, features, and labels of the data as *sample space*, *feature space*, and *label space*, respectively. Federated learning can be categorized into three types based on the distribution characteristics of the clients' data:

- *Horizontal (sample-based) federated learning*[111]: Data from different clients shares similar feature space but is very different in sample space. As an example, consider two hospitals in two different cities which collected similar information such as age, gender, and blood pressure about the individuals. In this case, the feature spaces are similar; but because the individuals who participated in the hospitals' data collections are from different cities, their intersection is most probably very small, and the sample spaces are hence very different.
- *Vertical (feature-based) federated learning*[111]: Clients' data are similar in sample space but very different in feature space. For example, two hospitals with different expertise in the same city might collect different information (different feature space) from almost the same individuals (similar sample space).
- *Hybrid federated learning*: Both feature space and sample space are different in the data from the clients. For example, consider a medical center with expertise in brain image analysis located in New York and a research center with expertise in protein research based in Berlin. Their data are completely different (image vs. protein data) and disjoint groups of individuals participated in the data collection of each center.

To illustrate the concept of federated learning, consider a scenario with two hospitals A and B. A and B possess lists X and Y, containing the age of their cancer patients, respectively. A simple federated mean algorithm to compute the average age of cancer patients in both hospitals without revealing the real values of X and Y works as follows: For the sake of brevity, we assume that both hospitals are selected in the first step and that the current global model parameters (average age) in the second step are zero (see federated learning steps).

- Hospital A computes the average age (M_X) and number of its cancer patients (N_X). Hospital B does the same,

resulting in M_Y, N_Y. Here, X and Y are private data while M_X, N_X, M_Y, N_Y are the parameters extracted from the private data.

- The server obtains the values of local model parameters from the hospitals and computes the global mean as follows:

$$M_G = \frac{(M_X \times N_X) + (M_Y \times N_Y)}{N_X + N_Y}$$

The emerging demand for federated learning gave rise to a wealth of both simulation[112,113] and production-oriented[114,115] open source frameworks. Additionally, there are AI platforms whose goal is to apply federated learning to real-world health care settings[116,117]. In the following, we survey studies on federated AI techniques in biomedicine and health care (►Table 2). Recent studies in this regard mainly focused on horizontal federated learning and there are a few vertical or hybrid federated learning algorithms applicable to genomic and biomedical data.

Several studies provided solutions for the lack of sufficient data due to the privacy challenges in the medical imaging domain.[117–123] For instance, Sheller et al developed a supervised DNN in a federated way for semantic segmentation of brain gliomas from magnetic resonance imaging scans.[118] Chang et al[123] simulated a distributed DNN in which multiple participants collaboratively update model weights using training heuristics such as single weight transfer and cyclical weight transfer (CWT). They evaluated this distributed model using image classification tasks on medical image datasets such as mammography and retinal fundus image collections, which were evenly distributed among the participants. Balachandar et al[122] optimized CWT for cases where the datasets are unevenly distributed across participants. They assessed their optimization methods on simulated diabetic retinopathy detection and chest radiograph classification.

Federated Cox regression, linear regression, logistic regression as well as Chi-square test have been developed for sensitive biomedical data that is vertically or horizontally distributed.[124–129] VERTICOX[124] is a vertical federated Cox regression model for survival analysis, which employs the alternating direction method of multiplier (ADMM) framework[130] and is evaluated on acquired immunodeficiency syndrome (AIDS) and breast cancer survival datasets. Similarly, WebDISCO[125] presents a federated Cox regression model but for horizontally distributed survival data. The grid binary logistic regression (GLORE)[126] and the expectation propagation logistic regression (EXPLORER)[127] implemented a horizontally federated logistic regression for medical data.

Unlike GLORE, EXPLORER supports asynchronous communication and online learning functionality so that the system can continue collaborating in case a participant is

Table 2 Literature for FL and hybrid approaches in biomedicine

Authors	Year	Technique	Model	Application
Sheller et al[118]	2018	FL	DNN	Medical image segmentation
Chang et al[123] Balachandar et al[122]	2018 2020	FL	Single weight transfer Cyclic weight transfer	Medical image segmentation
Nasirigerdeh et al[129]	2020	FL	Linear regression Logistic regression χ^2 statistics	Genetic associations
Wu et al[126] Wang et al[127] Li et al[128]	2012 2013 2016	FL	Logistic regression	Genetic associations
Dai et al[124] Lu et al[125]	2020 2015	FL	Cox regression	Survival analysis
Brisimi et al[132]	2018	FL	Support vector machines	Classifying electronic health records
Huang et al[133]	2018	FL	Adaptive boosting ensemble	Classifying medical data
Liu et al[134]	2018	FL	Autonomous deep learning	Classifying medical data
Chen et al[135]	2019	FL	Transfer learning	Training wearable health care devices
Li et al[150]	2020	FL + DP	DNN	Medical image segmentation
Li et al[149]	2019	FL + DP	Domain adoption	Medical image pattern recognition
Choudhury et al[159]	2019	FL + DP	Neural network Support vector machine Logistic regression	Classifying electronic health records
Constable et al[52]	2015	FL + SMPC	Statistical analysis (e.g., χ^2 statistics, ...)	Genetic associations
Lee et al[158]	2019	FL + HE	Context-specific hashing	Learning patient similarity
Kim et al[156]	2019	FL + DP + HE	Logistic regression	Classifying medical data

Abbreviations: DP, differential privacy; FL, federated learning; HE, homomorphic encryption; SMPC, secure multiparty computation.

absent or if communication is interrupted. Li et al presented VERTIGO,[128] a vertical grid logistic regression algorithm designed for vertically distributed biological datasets such as breast cancer genome and myocardial infarction data. Nasirigerdeh et al[129] developed a horizontally federated tool set for GWAS, called *sPLINK*, which supports Chi-square test, linear regression, and logistic regression. Notably, federated results from *sPLINK* on distributed datasets are the same as those from aggregated analysis conducted with *PLINK*.[131] Moreover, they showed that *sPLINK* is robust against heterogeneous (imbalanced) data distributions across clients and does not lose its accuracy in such scenarios.

There are also studies that combine federated learning with other traditional AI modeling techniques such as ensemble learning, support vector machines (SVMs), and principal component analysis (PCA).[132–136] Brisimi et al[132] presented a federated soft-margin support vector machine (sSVM) for distributed electronic health records. Huang et al[133] introduced LoAdaBoost, a federated adaptive boosting method for learning biomedical data such as intensive care unit data from distinct hospitals[137] while Liu et al[134] trained a federated autonomous deep learner to this end. There have also been a couple of attempts at incorporating federated learning into multitask learning and transfer learning in general.[138–140] However, to the best of our knowledge, FedHealth[135] is the only federated transfer learning framework specifically designed for health care applications. It enables users to train personalized models for their wearable health care devices by aggregating the data from different organizations without compromising privacy.

One of the major challenges for adopting federated learning in large scale genomics and biomedical applications is the significant network communication overhead, especially for complex AI models such as DNNs that contain millions of model parameters and require thousands of iterations to converge. A rich body of literature exists to tackle this challenge, known as communication-efficient federated learning.[141–144]

Another challenge in federated learning is the possible accuracy loss from the aggregation process if the data distribution across the clients is heterogeneous (i.e., not independent and identically distributed [IID]). More specifically, federated learning can deal with non-IID data while preserving the model accuracy if the learning model is simple such as ordinary least squares linear regression (*sPLINK*[129]). However, when it comes to learning complex models such as DNNs, the global model might not converge on non-IID data across the clients. Zhao et al[145] showed that simple averaging of the model parameters in the server significantly diminishes the accuracy of a convolutional neural network model in highly skewed non-IID settings. Developing the aggregation strategies which are robust against non-IID scenarios is still an open and interesting problem in federated learning.

Finally, federated learning is based on the assumption that the centralized server is honest and not compromised, which is not necessarily the case in real applications. To relax this assumption, differential privacy or cryptographic techniques can be leveraged in federated learning, which is covered in the next section. For further reading on further directions of federated learning in general, we refer the reader to comprehensive surveys.[110,146,147]

Hybrid Privacy-Preserving Techniques

The hybrid techniques combine federated learning with the other paradigms (cryptographic techniques and differential privacy) to enhance privacy or provide privacy guarantees (►Table 2). Federated learning preserves privacy to some extent because it does not require the health institutes to share the patients' data with the central server. However, the model parameters that participants share with the server might be abused to reveal the underlying private data if the coordinator is compromised.[148] To handle this issue, the participants can leverage differential privacy and add noise to the model parameters before sending them to the server (FL + DP)[149–153] or they employ HE (FL + HE),[55,154,155] SMPC (FL + SMPC) or both DP and HE (FL + DP + HE)[103,156,157] to securely share the parameters with the server.[51,158]

In the genomic and biomedical field, several hybrid approaches have been presented recently. Li et al[149] presented a federated deep learning framework for magnetic resonance brain image segmentation in which the client side provides differential privacy guarantees on selecting and sharing the local gradient weights with the server for imbalanced data. A recent study[150] extracted neural patterns from brain functional magnetic resonance images by developing a privacy-preserving pipeline that analyzes image data of patients having different psychiatric disorders using federated domain adaptation methods. Choudhury et al[159] developed a federated differential privacy mechanism for gradient-based classification on electronic health records.

There are also some studies that incorporate federate learning with cryptographic techniques. For instance, Constable et al[52] implemented a privacy-protecting structure for federated statistical analysis such as χ^2 statistics on GWAS while maintaining privacy using SMPC. In a slightly different approach, Lee et al[158] presented a privacy-preserving platform for learning patient similarity in multiple hospitals using a context-specific hashing approach which employs HE to limit the privacy leakage. Moreover, Kim et al[156] presented a privacy-preserving federated logistic regression algorithm for horizontally distributed diabetes and intensive care unit datasets. In this approach, the logistic regression ensures privacy by making the aggregated weights differentially private and encrypting the local weights using HE.

Incorporating HE, SMPC, and differential privacy into federated learning brings about enhanced privacy but it combines the limitations of the approaches, too. FL + HE puts much more computational overhead on the server, since it requires to perform aggregation on the encrypted model parameters from the clients. The network communication overhead is exacerbated in FL + SMPC, because clients need to securely share the model parameters with multiple

computing parties instead of one. FL + DP might result in inaccurate models because of adding noise to the model parameters in the clients.

Comparison

We compare the privacy-preserving techniques (HE, SMPC, differential privacy, federated learning, and the hybrid approaches) using various performance and privacy criteria such as computational/communication efficiency, accuracy, privacy guarantee, and exchanging sensitive traffic through network and privacy of exchanged traffic (►Fig. 3). We employ a generic ranking (lowest = 1 to highest = 6)[37] for all comparison criteria except for *privacy guarantee* and *exchanging sensitive traffic through network*, which are binary criteria. This comparison is made under the assumption of applying a complex model (e.g., DNN with a huge number of model parameters) on a large sensitive genomics dataset distributed across dozens of clients in IID configuration. Additionally, there are a few computing parties in SMPC (practical configuration).

Computational efficiency is an indicator of the extra computational overhead an approach incurs to preserve privacy. According to ►Fig. 3, federated learning is best from this perspective because it follows the paradigm of "bringing computation to data", distributing computational overhead among the clients. HE and SMPC are based on the paradigm of moving data to computation. In HE, encryption of the whole private data in the clients and carrying out computation on encrypted data by the computing party causes a huge amount of overhead. In SMPC, a couple of computing parties process the secret shares from dozens of clients, incurring considerable computational overhead. Among the hybrid approaches, FL + DP has the best computational efficiency given the lower overhead of the two approaches whereas FL + HE has the highest overhead because the aggregation process on encrypted parameters is computationally expensive.

Network communication efficiency indicates how efficient an approach utilizes the network bandwidth. The less data traffic is exchanged in the network, the more communication efficient is the approach. Federated learning is the least efficient approach from the communication aspect since exchanging a large number of model parameter values between the clients and the server generates a huge amount of network traffic. Notice that network bandwidth usage of federated learning is independent of the clients' data because federated learning does not move data to computation but depends on the model complexity (i.e., the number of model parameters). The next approach in this regard is SMPC, where not only each participant sends a large traffic (almost as big as its data) to each computing party but also each computing party exchanges intermediate results (which might be large) with the other computing parties through the network. Although recent research has shown that there is still potential for reducing the communication overhead in SMPC,[160] many limitations cannot be fully overcome. The network overhead of HE comes from sharing the encrypted data of the clients (assumed to be almost as big as the data itself) with the computing party, which is small compared with network traffic generated by federated learning and SMPC. The best approach is differential privacy with no network overhead. Accordingly, FL + DP and FL + SMPC are the best and worst among the hybrid approaches from a communication efficiency viewpoint, respectively.

Accuracy of the model in a privacy-preserving approach is a crucial factor in whether to adopt the approach. In the assumed configuration, SMPC and federated learning are the most accurate approaches incurring little accuracy loss in the final model. Next is differential privacy where the added noise can considerably affect the model accuracy. The worst approach is HE whose accuracy loss is due to approximating the non-linear operations using addition and multiplication (e.g., least squares approximation[65]). In the hybrid approaches, FL + SMPC is the best and FL + DP is the worst considering the accuracy of SMPC and differential privacy approaches.

The rest of the comparison measures are privacy related. The traffic transferred from the clients (participants) to the server (computing parties) is highly sensitive if it carries the private data of the clients. HE and SMPC send the encrypted form of the clients' private data to the server. Federated

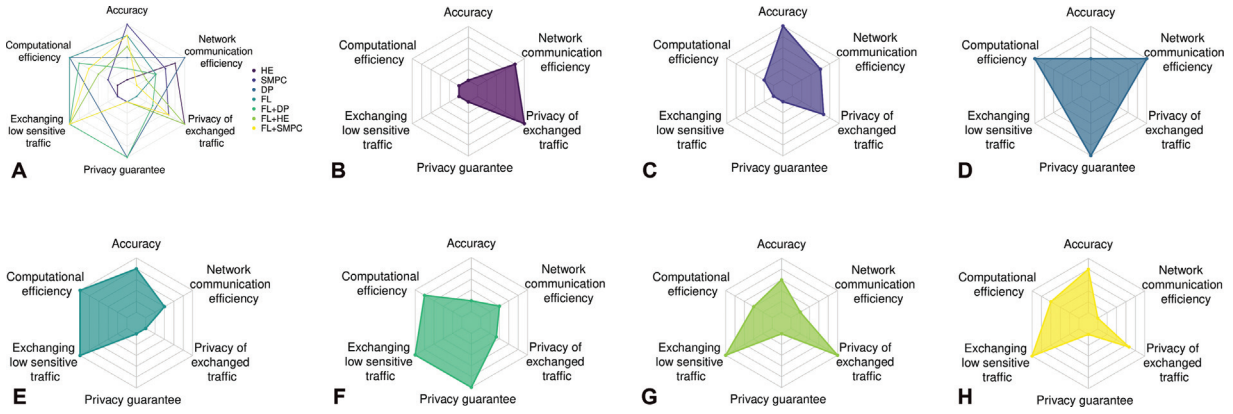

Fig. 3 Comparison radar plots for all (**A**) and each of (**B–H**) the privacy preserving approaches including homomorphic encryption (HE), secure multiparty computation (SMPC), differential privacy (DP), federated learning (FL) and hybrid techniques (FL + DP, FL + HE and FL + SMPC). (**A**) All. (**B**) HE. (**C**) SMPC. (**D**) DP. (**E**) FL. (**F**) FL + DP. (**G**) FL + HE. (**H**) FL + SMPC.

learning and hybrid approaches share only the model parameters with the server. In HE, if the server has the key to decrypt the traffic from the clients, the whole private data of the clients will be revealed. The same holds if the computing parties in SMPC collude with each other. This might or might not be the case for the other approaches (e.g., federated learning) depending on the exchanged model parameters and whether they can be abused to infer the underlying private data.

Privacy of the exchanged traffic indicates how much the traffic is kept private from the server. In HE/SMPC, the data are encrypted first and then shared with the server, which is reasonable since it is the clients' private data. In federated learning, the traffic (model parameters) is directly shared with the server assuming that it does not reveal any details regarding individual samples in the data. The aim of the hybrid approaches is to hide the real values of the model parameters from the server to minimize the possibility of inference attacks using the model parameters. FL + HE is the best among the hybrid approaches from this viewpoint.

Privacy guarantee is a metric which quantifies the degree to which the privacy of the clients' data can be preserved. Differential privacy and the corresponding hybrid approach (FL + DP) are the only approaches providing a privacy guarantee, whereas all other approaches can only protect the privacy under a set of certain assumptions. In HE, the server must not have the decryption key; in SMPC, not all computing parties must collude with each other; in federated learning, the model parameters should not give any detail about a sample in the clients' data.

Discussion and Open Problems

In HE, a single computing party carries out computation over the encrypted data from the clients. In SMPC, multiple computing parties perform operations on the secret shares from the clients. In federated learning, a single server aggregates the local model parameters shared by the clients. From a practical point of view, HE and SMPC that follow the paradigm of "move data to computation" do not scale as the number of clients or data size in clients become large. This is because they put the computational burden on a single or a few computing parties. Federated learning, on the other hand, distributes the computation across the clients (aggregation on the server is not computationally heavy) but the communication overhead between the server and clients is the major challenge to scalability of federated learning. The hybrid approaches inherit this issue and it is exacerbated in FL + SMPC. Combining HE with federated learning (FL + HE) adds another obstacle (computational overhead) to the scalability of federated learning. There is a growing body of literature on communication-efficient approaches to federated learning that can dramatically improve the scalability of federated learning and make it suitable for large-scale applications including those in biomedicine.

Given that federated learning is the most promising approach from the scalability viewpoint, it can be used as a standalone approach as long as inferring the clients' data from the model parameters is practically impossible. Otherwise, it should be combined with differential privacy to avoid possible inference attacks and exposure of clients' private data and to provide privacy guarantee. The accuracy of the model will be satisfactory in federated learning but it might be deteriorated in FL + DP. A realistic trade-off needs to be considered depending on the application of interest.

Moreover, differential privacy can have many practical applications in biomedicine as a standalone approach. It works very well for low-sensitivity queries such as counting queries (e.g., number of patients with a specific disease) on genomic databases and their generalizations (e.g., histograms) since the presence or absence of an individual changes the query's response by at most one. Moreover, it can be employed to release summary statistics of GWAS such as χ^2 and p-values in a differentially private manner while keeping the accuracy acceptable. A novel promising research direction is to incorporate differential privacy in deep generative models to generate synthetic genomic and biomedical data.

Future studies can investigate how to reach a compromise between scalability, privacy, and accuracy in real-world settings. The communication overhead of federated learning is still an open problem since although state-of-the-art approaches considerably reduce the network overhead, they adversely affect the accuracy of the model. Hence, novel approaches are required to preserve the accuracy, which is of great importance in biomedicine, while making federated learning communication efficient.

Adopting federated learning in non-IID settings, where genomic and biomedical datasets across different hospitals/medical centers are heterogeneous, is another important challenge to address. This is because typical aggregation procedures such as simple averaging do not work well for these settings, yielding inaccurate models. Hence, new aggregation procedures are required to tackle non-IID scenarios. Moreover, current communication-efficient approaches which were developed for an IID setting might not be applicable to heterogeneous scenarios. Consequently, new techniques are needed to reduce network overhead in these settings, while keeping the model accuracy satisfactory.

Combining differential privacy with federated learning to enhance privacy and to provide a privacy guarantee is still a challenging issue in the field. It becomes even more challenging for health care applications, where accuracy of the model is of crucial importance. Moreover, the concept of privacy guarantee in differential privacy has been defined for local settings. In distributed scenarios, a dataset might be employed multiple times to train different models with various privacy budgets. Therefore, a new formulation of privacy guarantee should be proposed for distributed settings.

Conclusion

For AI techniques to succeed, big biomedical data needs to be available and accessible. However, the more AI models are trained on sensitive biological data, the more the awareness

about the privacy issues increases, which, in turn, necessitate strategies for shielding the data.[70] Hence, privacy-enhancing techniques are crucial to allow AI to benefit from the sensitive biological data.

Cryptographic techniques, differential privacy, and federated learning can be considered as the prime strategies for protecting personal data privacy. These emerging techniques are based on either securing sensitive data, perturbing it or not moving it off site. In particular, cryptographic techniques securely share the data with a single (HE) or multiple computing parties (SMPC); differential privacy adds noise to sensitive data and quantifies privacy loss accordingly, while federated learning enables collaborative learning under orchestration of a centralized server without moving the private data outside local environments.

All of these techniques have their own strengths and limitations. HE and SMPC are more communication efficient compared with federated learning but they are computationally expensive since they move data to computation and put the computational burden on a server or a few computing parties. Federated learning, on the other hand, distributes computation across the clients but suffers from high network communication overhead. Differential privacy is an efficient approach from a computational and a communication perspective but it introduces accuracy loss by adding noise to data or model parameters. Hybrid approaches are studied to combine the advantages or to overcome the disadvantages of the individual techniques. We argued that federated learning as a standalone approach or in combination with differential privacy is the most promising approach to be adopted in biomedicine. We discussed the open problems and challenges in this regard including the balance of communication efficiency and model accuracy in non-IID settings, and the need for a new notion of privacy guarantee for distributed biomedical datasets.

Incorporating privacy into the analysis of genomic and biomedical data is still an open challenge, yet preliminary accomplishments are promising to bring practical privacy even closer to real-world settings. Future research should investigate how to achieve a trade-off between scalability, privacy, and accuracy in real biomedical applications.

Funding

The FeatureCloud project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No. 826078. This publication reflects only the authors' view and the European Commission is not responsible for any use that may be made of the information it contains. The work of J.B. and T. K. was also supported by the Horizon 2020 project REPOTRIAL (No. 777111). M.L., T.K., and J.B. have further been supported by BMBF project Sys_CARE (01ZX1908A). M.L. and J.B. were also supported by BMBF project SyMBoD (01ZX1910D). J.B.'s contribution was also supported by his VILLUM Young Investigator grant (nr. 13154).

Conflict of Interest

None declared.

Acknowledgments

►**Figs. 1** and **2** have been created with BioRender.com.

The authors would like to thank FeatureCloud Consortium members, Bela Bihari, Tobias Frisch, Anne Hartebrodt, Anne-Christin Hauschild, Dominik Heider, Andreas Holzinger, Walter Hotzendorfer, Markus Kastelitz, Rudolf Mayer, Cristian Nogales, Anastasia Pustozerova, Richard Rottger, Harald H.H.W. Schmidt, Ameli Schwalber, Christof Tschohl, and Andrea Wohner for their helpful comments toward improving the paper.

References

1 Schwarting W, Alonso-Mora J, Rus D. Planning and decision-making for autonomous vehicles. Annu Rev Control Robot Auton Syst 2018;1(01):187–210
2 Gehring J, Auli M, Grangier D, Yarats D, Dauphin YN. Convolutional sequence to sequence learning. Paper presented at: Proceedings of the 34th International Conference on Machine Learning. JMLR. Org. Volume 70:1243–12522017
3 Xiong W, Wu L, Alleva F, Droppo J, Huang X, Stolcke A. The Microsoft 2017 conversational speech recognition system. Paper presented at: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE; 2018:5934–5938
4 Holzinger A, Kieseberg P, Weippl E, Tjoa AM. Current Advances, Trends and Challenges of Machine Learning and Knowledge Extraction: From Machine Learning to Explainable AI in Springer Lecture Notes in Computer Science LNCS 11015. Cham: Springer; 2018:1–8
5 Gómez-Bombarelli R, Wei JN, Duvenaud D, et al. Automatic chemical design using a data-driven continuous representation of molecules. ACS Cent Sci 2018;4(02):268–276
6 Ma J, Yu MK, Fong S, et al. Using deep learning to model the hierarchical structure and function of a cell. Nat Methods 2018; 15(04):290–298
7 Nie D, Trullo R, Lian J, et al. Medical image synthesis with deep convolutional adversarial networks. IEEE Trans Biomed Eng 2018;65(12):2720–2730
8 Hosny A, Parmar C, Quackenbush J, Schwartz LH, Aerts HJWL. Artificial intelligence in radiology. Nat Rev Cancer 2018;18(08): 500–510
9 Beam AL, Kohane IS. Big data and machine learning in health care. JAMA 2018;319(13):1317–1318
10 Yu KH, Beam AL, Kohane IS. Artificial intelligence in healthcare. Nat Biomed Eng 2018;2(10):719–731
11 Yu MK, Ma J, Fisher J, Kreisberg JF, Raphael BJ, Ideker T. Visible machine learning for biomedicine. Cell 2018;173(07): 1562–1565
12 Chen H, Engkvist O, Wang Y, Olivecrona M, Blaschke T. The rise of deep learning in drug discovery. Drug Discov Today 2018;23(06): 1241–1250
13 Wainberg M, Merico D, Delong A, Frey BJ. Deep learning in biomedicine. Nat Biotechnol 2018;36(09):829–838
14 Min S, Lee B, Yoon S. Deep learning in bioinformatics. Brief Bioinform 2017;18(05):851–869
15 Litjens G, Kooi T, Bejnordi BE, et al. A survey on deep learning in medical image analysis. Med Image Anal 2017;42:60–88
16 Shen D, Wu G, Suk HI. Deep learning in medical image analysis. Annu Rev Biomed Eng 2017;19:221–248
17 Jiang F, Jiang Y, Zhi H, et al. Artificial intelligence in healthcare: past, present and future. Stroke Vasc Neurol 2017;2(04): 230–243
18 Libbrecht MW, Noble WS. Machine learning applications in genetics and genomics. Nat Rev Genet 2015;16(06):321–332
19 Nemati S, Holder A, Razmi F, Stanley MD, Clifford GD, Buchman TG. An interpretable machine learning model for accurate

prediction of sepsis in the ICU. Crit Care Med 2018;46(04): 547–553
20 Teare P, Fishman M, Benzaquen O, Toledano E, Elnekave E. Malignancy detection on mammography using dual deep convolutional neural networks and genetically discovered false color input enhancement. J Digit Imaging 2017;30(04):499–505
21 Veta M, van Diest PJ, Willems SM, et al. Assessment of algorithms for mitosis detection in breast cancer histopathology images. Med Image Anal 2015;20(01):237–248
22 Naveed M, Ayday E, Clayton EW, et al. Privacy in the genomic era. ACM Comput Surv 2015;48(01):1–44
23 Shokri R, Stronati M, Song C, Shmatikov V. Membership inference attacks against machine learning models. Paper presented at: 2017 IEEE Symposium on Security and Privacy (SP). IEEE; 2017:3–18
24 Papernot N, McDaniel P, Sinha A, Wellman MP. SoK: Security and privacy in machine learning. Paper presented at: 2018 IEEE European Symposium on Security and Privacy (EuroS&P). IEEE; 2018:399–414
25 Zhang C, Bengio S, Hardt M, Recht B, Vinyals O. Understanding deep learning requires rethinking generalization. Paper presented at: Proceedings of the International Conference on Learning Representations (ICLR); 2017
26 Zhang Y, Jia R, Pei H, Wang W, Li B, Song D. The secret revealer: generative model-inversion attacks against deep neural networks. Paper presented at: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2020:253–261
27 Shringarpure SS, Bustamante CD. Privacy risks from genomic data-sharing beacons. Am J Hum Genet 2015;97(05):631–646
28 Homer N, Szelinger S, Redman M, et al. Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. PLoS Genet 2008;4(08):e1000167
29 Harmanci A, Gerstein M. Analysis of sensitive information leakage in functional genomics signal profiles through genomic deletions. Nat Commun 2018;9(01):2453
30 Wang R, Li YF, Wang X, Tang H, Zhou X. Learning your identity and disease from research papers: information leaks in genome wide association study. Paper presented at: Proceedings of the 16th ACM Conference on Computer and Communications Security; 2009:534–544
31 Genomics GAGlobal Alliance for Genomics and Health. GENOMICS. A federated ecosystem for sharing genomic, clinical data. Science 2016;352(6291):1278–1280
32 Humbert M, Ayday E, Hubaux J-P, Telenti A. Addressing the concerns of the lacks family: quantification of kin genomic privacy. Paper presented at: Proceedings of the 2013 ACM SIGSAC Conference on Computer& Communications Security; 2013:1141–1152
33 Zerhouni EA, Nabel EG. Protecting aggregate genomic data. Science 2008;322(5898):44–44
34 Erlich Y, Narayanan A. Routes for breaching and protecting genetic privacy. Nat Rev Genet 2014;15(06):409–421
35 General Data Protection Regulation(GDPR). Accessed May 15, 2020 at: https://gdpr-info.eu/2020
36 Cohen A, Nissim K. Towards formalizing the GDPR's notion of singling out. Paper presented at: Proceedings of the National Academy of Sciences; 2020
37 Aziz MMA, Sadat MN, Alhadidi D, et al. Privacy-preserving techniques of genomic data—a survey. Brief Bioinform 2019;20 (03):887–895
38 Xu J, Glicksberg BS, Su C, Walker P, Bian J, Wang F. Federated learning for healthcare informatics. J Healthc Inform Res 2020;5 (01):1–19
39 Kaissis GA, Makowski MR, Rückert D, Braren RF. Secure, privacy-preserving and federated machine learning in medical imaging. Nat Mach Intell 2021;3:474–484
40 Chen F, Wang S, Jiang X, et al. PRINCESS: privacy-protecting rare disease international network collaboration via encryption through software guard extensionS. Bioinformatics 2017;33 (06):871–878
41 Chen F, et al. Premix: privacy-preserving estimation of individual admixture. Paper presented at: AMIA Annual Symposium Proceedings, vol. 2016. American Medical Informatics Association; 2016:1747
42 Fuhry B, et al. Hardidx: Practical and Secure Index with sgx. Paper presented at: IFIP Annual Conference on Data and Applications Security and Privacy. Springer; 2017:386–408
43 Kaplan D, Powell J, Woller T. AMD memory encryption. White paper; Accessed April, 2016 at: http://developer.amd.com/wordpress/media/2013/12/AMD_Memory_Encryption_Whitepaper_v7-Public.pdf
44 Cho H, Wu DJ, Berger B. Secure genome-wide association analysis using multiparty computation. Nat Biotechnol 2018;36(06): 547–551
45 Bonte C, Makri E, Ardeshirdavani A, Simm J, Moreau Y, Vercauteren F. Towards practical privacy-preserving genome-wide association study. BMC Bioinformatics 2018;19(01):537
46 Jagadeesh KA, Wu DJ, Birgmeier JA, Boneh D, Bejerano G. Keeping patient phenotypes and genotypes private while seeking disease diagnoses. bioRxiv 2019:746230
47 Kim M, Lauter K. Private Genome Analysis through Homomorphic Encryption. In: BMC Medical Informatics and Decision Making; 15:S3 Springer; 2015
48 Lauter K, López-Alt A, Naehrig M. Private Computation on Encrypted Genomic Data. Paper presented at: International Conference on Cryptology and Information Security in Latin America. Springer; 2014:3–27
49 Lu WJ, Yamada Y, Sakuma J. Privacy-Preserving Genome-Wide Association Studies on Cloud Environment Using Fully Homomorphic Encryption. In: BMC Medical Informatics and Decision Making;15:S1 Springer; 2015
50 Zhang Y, Dai W, Jiang X, Xiong H, Wang S. Foresee: Fully Outsourced Secure Genome Study Based on Homomorphic Encryption. In: BMC Medical Informatics and Decision Making;15:S5 Springer; 2015
51 Kamm L, Bogdanov D, Laur S, Vilo J. A new way to protect privacy in large-scale genome-wide association studies. Bioinformatics 2013;29(07):886–893
52 Constable SD, Tang Y, Wang S, Jiang X, Chapin S. Privacy-preserving GWAS analysis on federated genomic datasets. In: BMC Medical Informatics and Decision Making; 15:S2 Springer; 2015
53 Zhang Y, Blanton M, Almashaqbeh G. Secure distributed genome analysis for GWAS and sequence comparison computation. In: BMC Medical Informatics and Decision Making; 15:S4 BioMed Central; 2015
54 Mohassel P, Zhang Y. Secureml: a system for scalable privacy-preserving machine learning. Paper presented at: 2017 IEEE Symposium on Security and Privacy (SP). IEEE; 2017:19–38
55 Hasan Z, Mahdi MSR, Mohammed N. Secure count query on encrypted genomic data: a survey. IEEE Internet Comput 2018; 22(02):71–82
56 Sadat MN, Al Aziz MM, Mohammed N, Chen F, Jiang X, Wang S. SAFETY: Secure gwAs in Federated Environment through a hYbrid Solution. IEEE/ACM Trans Comput Biol Bioinformatics 2019;16(01):93–102
57 Gentry C. Fully homomorphic encryption using ideal lattices Paper presented at: Proceedings of the 41st Annual ACM Symposium on Theory of Computing. 2009:169–178
58 Paillier P. Public-key cryptosystems based on composite degree residuosity classes. Paper presented at: International Conference on the Theory and Applications of Cryptographic Techniques; 1999:223–238
59 Gennaro R, Rabin MO, Rabin T. Simplified vss and Fast-Track Multiparty Computations with Applications to Threshold Cryptography. Paper presented at: Proceedings of the 17th annual

ACM Symposium on Principles of Distributed Computing; 1998: 101–111
60 Cramer R, Damgård IB, Nielsen JB. Secure Multiparty Computation. Cambridge University Press; 2015
61 Shamir A. How to share a secret. Commun ACM 1979;22(11): 612–613
62 Rabin MO. How to exchange secrets with oblivious transfer. IACR Cryptol. ePrint Arch 2005 2005(187)
63 Yao AC-C. How to generate and exchange secrets. Paper presented at: 27th Annual Symposium on Foundations of Computer Science (sfcs 1986). IEEE; 1986:162–167
64 Boneh D, Boyen X, Halevi S. Chosen ciphertext secure public key threshold encryption without random oracles. Paper presented at: Cryptographers' Track at the RSA Conference. Springer; 2006: 226–243
65 Kim M, Song Y, Wang S, Xia Y, Jiang X. Secure logistic regression based on homomorphic encryption: design and evaluation. JMIR Med Inform 2018;6(02):e19
66 Morshed T, Alhadidi D, Mohammed N. Parallel linear regression on encrypted data. Paper presented at: 2018 16th Annual Conference on Privacy, Security and Trust (PST). IEEE; 2018:1–5
67 Shi H, Jiang C, Dai W, et al. Secure multi-pArty computation grid LOgistic REgression (SMAC-GLORE). BMC Med Inform Decis Mak 2016;16(03, Suppl 3):89
68 Bloom JM. Secure multi-party linear regression at plaintext speed. arXiv preprint arXiv:1901.09531; 2019
69 Mittos A, Malin B, De Cristofaro E. Systematizing genome privacy research: a privacy-enhancing technologies perspective. Proc Privacy Enhancing Technol 2019;2019(01):87–107
70 Berger B, Cho H. Emerging technologies towards enhancing privacy in genomic data sharing. Genome Biol 2019;20(01):128
71 Chialva D, Dooms A. Conditionals in homomorphic encryption and machine learning applications. arXiv preprint arXiv:1810.12380 2018
72 Alexandru AB, Pappas GJ. Secure Multi-party Computation for Cloud-Based Control. In: Privacy in Dynamical Systems 2020 (pp. 179-207). Singapore: Springer
73 Dwork C, McSherry F, Nissim K, Smith A. Calibrating noise to sensitivity in private data analysis. J Privacy Confidentiality 2016;7(03):17–51
74 Dwork C, Kenthapadi K, McSherry F, Mironov I, Naor M. Our data, ourselves: privacy via distributed noise generation. Paper presented at: Annual International Conference on the Theory and Applications of Cryptographic Techniques. Springer; 2006: 486–503
75 Nissim K, Steinke T, Wood A, et al. Differential privacy: a primer for a non-technical audience. Paper presented at: Privacy Law Scholars Conf; 2017
76 Erlingsson Ú, Pihur V, Korolova A. Rappor: Randomized aggregatable privacy-preserving ordinal response. Paper presented at: Proceedings of the 2014 ACM SIGSAC conference on computer and communications security; 2014:1054–1067
77 Thakurta AG, Vyrros AH, Vaishampayan US, et al. Learning new words. 2017. US Patent 9,594,741
78 Johnson A, Shmatikov V. Privacy-preserving data exploration in genome-wide association studies. Paper presented at: Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining; 2013:1079–1087
79 Aziz MMA, Ghasemi R, Waliullah M, Mohammed N. Aftermath of bustamante attack on genomic beacon service. BMC Med Genomics 2017;10(02, Suppl 2):43
80 Han Z, Lu L, Liu H. A differential privacy preserving approach for logistic regression in genome-wide association studies. Paper presented at: 2019 International Conference on Networking and Network Applications (NaNA). IEEE; 2019:181–185
81 Yu F, Rybar M, Uhler C, Fienberg SE. Differentially-private logistic regression for detecting multiple-SNP association in GWAS databases. Paper presented at: International Conference on Privacy in Statistical Databases. Springer; 2014:170–184
82 Honkela A, Das M, Nieminen A, Dikmen O, Kaski S. Efficient differentially private learning improves drug sensitivity prediction. Biol Direct 2018;13(01):1
83 Simmons S, Sahinalp C, Berger B. Enabling privacy-preserving GWASs in heterogeneous human populations. Cell Syst 2016;3 (01):54–61
84 Simmons S, Berger B. Realizing privacy preserving genome-wide association studies. Bioinformatics 2016;32(09):1293–1300
85 Fienberg SE, Slavkovic A, Uhler C. Privacy preserving GWAS data sharing. Paper presented at: 2011 IEEE 11th International Conference on Data Mining Workshops. IEEE; 2011:628–635
86 Uhlerop C, Slavković A, Fienberg SE. Privacy-preserving data sharing for genome-wide association studies. J Priv Confid 2013; 5(01):137–166
87 Yu F, Ji Z. Scalable privacy-preserving data sharing methodology for genome-wide association studies: an application to iDASH healthcare privacy protection challenge. BMC Med Inform Decis Mak 2014;14(01, Suppl 1):S3
88 Wang S, Mohammed N, Chen R. Differentially private genome data dissemination through top-down specialization. BMC Med Inform Decis Mak 2014;14(01, Suppl 1):S2
89 Beaulieu-Jones BK, Yuan W, Finlayson SG, Wu ZS. Privacy-preserving distributed deep learning for clinical data. arXiv preprint arXiv: 1812.01484 2018
90 Han Z, Liu H, Wu Z. A differential privacy preserving framework with nash equilibrium in genome-wide association studies. Paper presented at: 2018 International Conference on Networking and Network Applications (NaNA). IEEE; 2018:91–96
91 Tramèr F, Huang Z, Hubaux JP, Ayday E. Differential privacy with bounded priors: reconciling utility and privacy in genome-wide association studies. Paper presented at: Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security; 2015:1286–1297
92 Vu D, Slavkovic A. Differential privacy for clinical trial data: preliminary evaluations. Paper presented at: 2009 IEEE International Conference on Data Mining Workshops. IEEE; 2009: 138–143
93 Wan Z, Vorobeychik Y, Kantarcioglu M, Malin B. Controlling the signal: practical privacy protection of genomic data sharing through Beacon services. BMC Med Genomics 2017;10(02, Suppl 2):39
94 Kairouz P, Oh S, Viswanath P. The composition theorem for differential privacy. IEEE Trans Inf Theory 2017;63(06): 4037–4049
95 Cho H, Simmons S, Kim R, Berger B. Privacy-preserving biomedical database queries with optimal privacy-utility trade-offs. Cell Syst 2020;10(05):408–416.e9
96 Ji Z, Jiang X, Wang S, Xiong L, Ohno-Machado L. Differentially private distributed logistic regression using private and public data. BMC Med Genomics 2014;7(Suppl 1):S14
97 Abay NC, Zhou Y, Kantarcioglu M, Thuraisingham B, Sweeney L. Privacy preserving synthetic data release using deep learning. Paper presented at: Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Springer; 2018:510–526
98 Beaulieu-Jones BK, Wu ZS, Williams C, et al. Privacy-preserving generative deep neural networks support clinical data sharing. Circ Cardiovasc Qual Outcomes 2019;12(07):e005122
99 Jordon J, Yoon J, Van Der Schaar M. PATE-GAN: Generating synthetic data with differential privacy guarantees. In International conference on learning representations 2018 Sep 27
100 Fiume M, Cupak M, Keenan S, et al. Federated discovery and sharing of genomic data using Beacons. Nat Biotechnol 2019;37 (03):220–224
101 Raisaro JL, Gwangbae Choi, Pradervand S, et al. Protecting privacy and security of genomic data in I2B2 with homomorphic

encryption and differential privacy. IEEE/ACM Trans Comput Biol Bioinformatics 2018;15(05):1413–1426
102 Hardt M, Ligett K, McSherry F. A simple and practical algorithm for differentially private data release. In Advances in Neural Information Processing Systems, pp. 2339–2347, 2012
103 Raisaro JL, Troncoso-Pastoriza JR, Misbach M, et al. MedCo: enabling secure and privacy-preserving exploration of distributed clinical and genomic data. IEEE/ACM Trans Comput Biol Bioinformatics 2019;16(04):1328–1341
104 Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet 2006;38(08): 904–909
105 Yang J, Zaitlen NA, Goddard ME, Visscher PM, Price AL. Advantages and pitfalls in the application of mixed-model association methods. Nat Genet 2014;46(02):100–106
106 Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial networks. Communications of the ACM 2020 22;63(11): 139–144
107 Wang S, Jiang X, Singh S, et al. Genome privacy: challenges, technical approaches to mitigate risk, and ethical considerations in the United States. Ann N Y Acad Sci 2017;1387(01):73–83
108 Kieseberg P, Hobel H, Schrittwieser S, Weippl E, Holzinger A. Protecting anonymity in data-driven biomedical science. In:In: Holzinger A, Jurisica I, eds. Interactive Knowledge Discovery and Data Mining in Biomedical Informatics, Lecture Notes in Computer Science, LNCS 8401. Berlin Heidelberg: Springer; 2014:301–316
109 McMahan B, Moore E, Ramage D, Hampson S, Arcas BAy. Communication-efficient learning of deep networks from decentralized data. In Artificial Intelligence and Statistics. PMLR 2017; 54:1273–1282
110 Kairouz P, McMahan HB, Avent B, et al. Advances and open problems in federated learning. arXiv preprint arXiv:1912.04977 2019
111 Yang Q, Liu Y, Chen T, Tong Y. Federated machine learning: concept and applications. ACM Trans Intell Syst Technol 2019; 10(02):1–19
112 The TFF Authors. Tensor flow federated. 2019. Accessed June 18 2020 at: https://www.tensorflow.org/federated
113 Ryffel T, Trask A, Dahl M, et al. A generic framework for privacy preserving deep learning. arXiv preprint arXiv:1811.04017 2018
114 The FATE Authors. Federated AI technology enabler. 2019. Accessed June 18 2020 at: https://www.fedai.org/
115 The PaddleFL Authors. PaddleFL, 2019. Accessed June 18 2020 at: https://github.com/PaddlePaddle/PaddleFL
116 The FeatureCloud Authors. FeatureCloud, 2019. Accessed June 18 2020 at: https://featurecloud.eu/
117 Clara NVIDIA. The Clara training framework authors. 2019. Accessed June 18 2020 at: https://developer.nvidia.com/clara-medical-imaging
118 Sheller MJ, Reina GA, Edwards B, Martin J, Bakas S. Multi-institutional deep learning modeling without sharing patient data: a feasibility study on brain tumor segmentation. Paper presented at: International MICCAI Brainlesion Workshop. Springer; 2018:92–104
119 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 2018
120 Vepakomma P, Gupta O, Dubey A, Raskar R. Reducing leakage in distributed deep learning for sensitive health data. arXiv preprint arXiv:1812.00564 2019
121 Poirot MG, Vepakomma P, Chang K, Kalpathy-Cramer J, Gupta R, Raskar R. Split Learning for collaborative deep learning in healthcare. arXiv preprint arXiv:1912.12115 2019
122 Balachandar N, Chang K, Kalpathy-Cramer J, Rubin DL. Accounting for data variability in multi-institutional distributed deep learning for medical imaging. J Am Med Inform Assoc 2020;27 (05):700–708
123 Chang K, Balachandar N, Lam C, et al. Distributed deep learning networks among institutions for medical imaging. J Am Med Inform Assoc 2018;25(08):945–954
124 Dai W, Jiang X, Bonomi L, Li Y, Xiong H, Ohno-Machado L. VERTICOX: Vertically Distributed Cox Proportional Hazards Model Using the Alternating Direction Method of Multipliers. IEEE Transactions on Knowledge and Data Engineering 2020 Apr 22
125 Lu CL, Wang S, Ji Z, et al. WebDISCO: a web service for distributed cox model learning without patient-level data sharing. J Am Med Inform Assoc 2015;22(06):1212–1219
126 Wu Y, Jiang X, Kim J, Ohno-Machado L. Grid Binary LOgistic REgression (GLORE): building shared models without sharing data. J Am Med Inform Assoc 2012;19(05):758–764
127 Wang S, Jiang X, Wu Y, Cui L, Cheng S, Ohno-Machado L. EXpectation Propagation LOgistic REgRession (EXPLORER): distributed privacy-preserving online model learning. J Biomed Inform 2013;46(03):480–496
128 Li Y, Jiang X, Wang S, Xiong H, Ohno-Machado L. VERTIcal Grid lOgistic regression (VERTIGO). J Am Med Inform Assoc 2016;23 (03):570–579
129 Nasirigerdeh R, Torkzadehmahani R, Matschinske J, Frisch T, List M, Späth J, Wei ß S, Völker U, Heider D, Wenke NK, Kacprowski T. sPLINK: a federated, privacy-preserving tool as a robust alternative to meta-analysis in genome-wide association studies. BioRxiv 2020 Jan 1
130 Gabay D. Applications of the method of multipliers to variational inequalities. In: Studies in Mathematics and Its Applications. Elsevier 1983;15:299–331
131 Purcell S, Neale B, Todd-Brown K, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 2007;81(03):559–575
132 Brisimi TS, Chen R, Mela T, Olshevsky A, Paschalidis IC, Shi W. Federated learning of predictive models from federated electronic health records. Int J Med Inform 2018;112:59–67
133 Huang L, Yin Y, Fu Z, Zhang S, Deng H, Liu D. Loadaboost: loss-based adaboost federated machine learning on medical data. arXiv preprint arXiv: 1811.12629 2018
134 Liu D, Miller T, Sayeed R, Mandl KD. Fadl: federated-autonomous deep learning for distributed electronic health record. arXiv preprint arXiv:1811.11400 2018
135 Chen Y, Wang J, Yu C, Gao W, Qin X. FedHealth: A Federated Transfer Learning Framework for Wearable Healthcare. arXiv preprint arXiv:1907.09173 2019
136 Silva S, Gutman BA, Romero E, Thompson PM, Altmann A, Lorenzi M. Federated learning in distributed medical databases: meta-analysis of large-scale subcortical brain data. Paper presented at: 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019). IEEE; 2019
137 Pollard TJ, Johnson AEW, Raffa JD, Celi LA, Mark RG, Badawi O. The eICU Collaborative Research Database, a freely available multi-center database for critical care research. Sci Data 2018; 5:180178
138 Smith V, Chiang CK, Sanjabi M, Talwalkar AS. Federated multi-task learning. In: Advances in Neural Information Processing Systems2017:4424–4434
139 Corinzia L, Buhmann JM. Variational federated multi-task learning. arXiv preprint arXiv:1906.06268; 2019
140 Liu Y, Kang Y, Xing C, Chen T, Yang Q. A secure federated transfer learning framework. IEEE Intell Syst 2020;35(04):70–82
141 Gupta S, Agrawal A, Gopalakrishnan K, Narayanan P. Deep learning with limited numerical precision. Paper presented at: International Conference on Machine Learning; 2015: 1737–1746
142 Aji AF, Heafield K. Sparse Communication for Distributed Gradient Descent. Paper presented at: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing; 2017:440–445

143 McMahan HB, Moore E, Ramage D, Hampson S. Communication-efficient learning of deep networks from decentralized data. arXiv preprint arXiv: 1602.05629 2016
144 Tang Z, Shi S, Chu X, Wang W, Li B. Communication-efficient distributed deep learning: a comprehensive survey. arXiv preprint arXiv:2003.06307 2020
145 Zhao Y, Li M, Lai L, Suda N, Civin D, Chandra V. Federated learning with non-iid data. arXiv preprint arXiv:1806.00582 2018
146 Li Q, Wen Z, He B. Federated learning systems: vision, hype and reality for data privacy and protection. arXiv preprint arXiv:1907.09693 2019
147 Rieke N, Hancox J, Li W, et al. The future of digital health with federated learning. NPJ Digit Med 2020;3(01):119
148 Melis L, Song C, De Cristofaro E, Shmatikov V. Exploiting unintended feature leakage in collaborative learning. Paper presented at: 2019 IEEE Symposium on Security and Privacy (SP). IEEE; 2019:691–706
149 Li W, Milletarì F, Xu D, et al. Privacy-preserving federated brain tumour segmentation. Paper presented at: International Workshop on Machine Learning in Medical Imaging. Springer; 2019: 133–141
150 Li X, Gu Y, Dvornek N, Staib L, Ventola P, Duncan JS. Multi-site fMRI analysis using privacy-preserving federated learning and domain adaptation: ABIDE results. arXiv preprint arXiv:2001. 05647 2020
151 Geyer RC, Klein T, Nabi M. Differentially private federated learning: a client level perspective. arXiv preprint arXiv:1712. 07557 2017
152 Truex S, Baracaldo N, Anwar A, et al. A hybrid approach to privacy-preserving federated learning. Paper presented at: Proceedings of the 12th ACM Workshop on Artificial Intelligence and Security; 2019:1–11
153 Wei K, Li J, Ding M, et al. Federated learning with differential privacy: algorithms and performance analysis. IEEE Trans Inf Forensics Security 2020;15:3454–3469
154 Hardy S, Henecka W, Ivey-Law H, et al. Private federated learning on vertically partitioned data via entity resolution and additively homomorphic encryption. arXiv preprint arXiv:1711.10677 2017
155 Zhang C, Li S, Xia J, Wang W, Yan F, Liu Y. BatchCrypt: efficient homomorphic encryption for cross-silo federated learning. Paper presented at: 2020 {USENIX} Annual Technical Conference ({USENIX}{ATC} 20) 493–506
156 Kim M, Lee J, Ohno-Machado L, Jiang X. Secure and differentially private logistic regression for horizontally distributed data. IEEE Trans Inf Forensics Security 2019;15:695–710
157 Froelicher D, Egger P, Sousa JS, et al. UnLynx: a decentralized system for privacy-conscious data sharing. Proceedings on Privacy Enhancing Technologies. 2017;2017(04):232–250
158 Lee J, Sun J, Wang F, Wang S, Jun CH, Jiang X. Privacy-preserving patient similarity learning in a federated environment: development and analysis. JMIR Med Inform 2018;6(02):e20
159 Choudhury O, Gkoulalas-Divanis A, Salonidis T, et al. Differential privacy-enabled federated learning for sensitive health data. arXiv preprint arXiv:1910.02578 2019
160 Dankar FK, Madathil N, Dankar SK, Boughorbel S. Privacy-preserving analysis of distributed biomedical data: designing efficient and secure multiparty computations using distributed statistical learning theory. JMIR Med Inform 2019;7(02): e12702