Privacy-Preserving Artificial Intelligence Techniques in Biomedicine

Reihaneh Torkzadehmahani1 Reza Nasirigerdeh1,2 Markus List6 Julian Matschinske7,8 Julian Spaeth7,8 David B. Blumenthal3 Tim Kacprowski4,5 Nina Kerstin Wenke7,8 Jan Baumbach7,8,9

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

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

Methods Inf Med

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.

ISSN 0026-1270.

© 2022. The Author(s).
This is an open access article published by Thieme under the terms of the Creative Commons Attribution-NonDerivative-NonCommercial-License, permitting copying and reproduction so long as the original work is given appropriate credit. Contents may not be used for commercial purposes, or adapted, remixed, transformed or built upon. (https://creativecommons.org/licenses/by-nc-nd/4.0/)

Georg Thieme Verlag KG, Rüdigerstraße 14, 70469 Stuttgart, Germany
**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, text translation, and voice assistance. 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.

Success stories such as acquiring the compressed representation of drug-like molecules, modeling the hierarchical structure and function of a cell, and translating magnetic resonance images to computed tomography 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 but has also surpassed humans in applications such as sepsis prediction, malignancy detection in mammography, and mitosis detection in breast cancer.

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, several studies showed that AI techniques often do not maintain data privacy. For example, attacks known as membership inference can be used to infer an individual’s membership by querying over the dataset or the trained model, or by having access to certain statistics about the dataset. Homer et al. 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, 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. 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.

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. 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. 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. Aziz et al. investigated previous studies which employed differential privacy and cryptographic techniques for human genomic data. Kassis et al. briefly reviewed federated learning, differential privacy and cryptographic techniques applied in medical imaging. Xu et al. 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 Kassis et al., this paper covers a broader set of privacy preserving techniques including federated learning and hybrid approaches. In contrast with Xu et al., 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 and AMD memory encryption, 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 HE57–59 or SMPC60. 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 and multiparty computation, consider a scenario with two participants P1 and P2 and two computing parties C1 and C2. P1 and P2 possess the private data X and Y, respectively. The aim is to compute X + Y, where neither P1 nor P2 reveals its data to the computing parties. To this end, P1 and P2 generate random numbers RX and RV, respectively; P1 reveals RX to C1 and \((X - RX)\) to C2; likewise, P2 shares RV with C1 and Y – RV with C2. C2 computes \(RX + RV\) and sends it to C2 and C2 calculates \((X - RX) + (Y - RV)\) and reveals it to C1. Both C1 and C2 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 P1 and P2.

Notice that to preserve the data privacy, the computing parties C1 and C2 must be non-colluding. That is, C1 must not send RX and RV to C2 and C2 must not share \((X - RX)\) and \((Y - RV)\) with C1. 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 and garbled circuit, which is a two-party computation protocol in which each of the parties holds 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. 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 presented HE-based secure logistic and linear regression algorithms for medical data, respectively. Zhang et al., Kamm et al., and Shi et al. implemented a privacy-preserving logistic regression and Bloom proposed a secure linear regression based on SMPC for GWAS data. Cho 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), 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. For more details, we refer to the comprehensive review by Mittos et al.

Despite the promises of HE/SMPC-based privacy-preserving algorithms, the road for the wide adoption of HE/SMPC-based algorithms in genomics and biomedicine is long. The major limitations of HE are few supported operations and computational overhead. 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, 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. 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 and Apple 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.

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
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 \((\varepsilon, \delta)\)-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, \( \varepsilon \) and \( \delta \) are privacy loss parameters where lower values imply stronger privacy guarantees. \( \delta \) is an exceedingly

### Table 1: Literature for cryptographic techniques and differential privacy in biomedicine

<table>
<thead>
<tr>
<th>Authors</th>
<th>Year</th>
<th>Technique</th>
<th>Model</th>
<th>Application</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kim and Lauter</td>
<td>2015</td>
<td>HE</td>
<td>( \chi^2 ) statistics</td>
<td>Genetic associations</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Minor allele frequency</td>
<td>DNA comparison</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Hamming Distance</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Edit distance</td>
<td></td>
</tr>
<tr>
<td>Lu et al</td>
<td>2015</td>
<td>HE</td>
<td>( \chi^2 ) statistics</td>
<td>Genetic associations</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>( D' ) measure</td>
<td></td>
</tr>
<tr>
<td>Lauter et al</td>
<td>2014</td>
<td>HE</td>
<td>( D' ) and ( r^2 ) measure</td>
<td>Genetic associations</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Pearson goodness-of-fit</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>expectation maximization</td>
<td></td>
</tr>
<tr>
<td>Kim et al</td>
<td>2018</td>
<td>HE</td>
<td>Logistic regression</td>
<td>Medical decision-making</td>
</tr>
<tr>
<td>Morshed et al</td>
<td>2018</td>
<td>HE</td>
<td>Linear regression</td>
<td>Medical decision-making</td>
</tr>
<tr>
<td>Kamm et al</td>
<td>2013</td>
<td>SMPC</td>
<td>( \chi^2 ) statistics</td>
<td>Genetic associations</td>
</tr>
<tr>
<td>Constable et al</td>
<td>2015</td>
<td>SMPC</td>
<td>( \chi^2 ) statistics</td>
<td>Genetic associations</td>
</tr>
<tr>
<td>Zhang et al</td>
<td>2015</td>
<td>SMPC</td>
<td>Minor allele frequency</td>
<td>Genetic associations</td>
</tr>
<tr>
<td>Shi et al</td>
<td>2016</td>
<td>SMPC</td>
<td>Logistic regression</td>
<td>Genetic associations</td>
</tr>
<tr>
<td>Bloom</td>
<td>2019</td>
<td>SMPC</td>
<td>Linear regression</td>
<td>Genetic associations</td>
</tr>
<tr>
<td>Cho et al</td>
<td>2018</td>
<td>SMPC</td>
<td>Quality Control Population strati</td>
<td>Genetic associations</td>
</tr>
<tr>
<td>Johnson and Shmatikov</td>
<td>2013</td>
<td>DP</td>
<td>Distance-score mechanism</td>
<td>Querying genomic databases</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>( p )-value and ( \chi^2 ) statistics</td>
<td></td>
</tr>
<tr>
<td>Cho et al</td>
<td>2020</td>
<td>DP</td>
<td>( \alpha )-geometric mechanism</td>
<td>Querying biomedical databases</td>
</tr>
<tr>
<td>Aziz et al</td>
<td>2017</td>
<td>DP</td>
<td>Eliminating random positions</td>
<td>Querying genomic databases</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Biased random response</td>
<td></td>
</tr>
<tr>
<td>Han et al</td>
<td>2019</td>
<td>DP</td>
<td>Logistic regression</td>
<td>Genetic associations</td>
</tr>
<tr>
<td>Yu et al</td>
<td>2014</td>
<td>DP</td>
<td>Bayesian linear regression</td>
<td>Drug sensitivity prediction</td>
</tr>
<tr>
<td>Honkela et al</td>
<td>2018</td>
<td>DP</td>
<td>Bayesian linear regression</td>
<td>Drug sensitivity prediction</td>
</tr>
<tr>
<td>Simmons et al</td>
<td>2016</td>
<td>DP</td>
<td>EIGENSTRAT Linear mixed model</td>
<td>Genetic associations</td>
</tr>
<tr>
<td>Simmons and Berger</td>
<td>2016</td>
<td>DP</td>
<td>Nearest neighbor optimization</td>
<td>Genetic associations</td>
</tr>
<tr>
<td>Fienberg et al</td>
<td>2011</td>
<td>DP</td>
<td>Statistics such as ( p )-value</td>
<td>Genetic associations</td>
</tr>
<tr>
<td></td>
<td>2013</td>
<td></td>
<td>( \chi^2 ) and contingency table</td>
<td></td>
</tr>
<tr>
<td>Ulher et al</td>
<td>2014</td>
<td>DP</td>
<td>Statistics such as ( p )-value</td>
<td>Genetic associations</td>
</tr>
<tr>
<td>Yu and Ji</td>
<td>2014</td>
<td>DP</td>
<td>Statistics such as ( p )-value</td>
<td>Genetic associations</td>
</tr>
<tr>
<td>Wang et al</td>
<td>2014</td>
<td>DP</td>
<td>Statistics such as ( p )-value</td>
<td>Genetic associations</td>
</tr>
<tr>
<td>Abay et al</td>
<td>2018</td>
<td>DP</td>
<td>Deep autoencoder</td>
<td>Generating artificial biomedical data</td>
</tr>
<tr>
<td>Beaulieu et al</td>
<td>2019</td>
<td>DP</td>
<td>GAN</td>
<td>Simulating SPRINT trial</td>
</tr>
<tr>
<td>Jordon et al</td>
<td>2018</td>
<td>DP</td>
<td>GAN</td>
<td>Generating artificial biomedical data</td>
</tr>
</tbody>
</table>

Abbreviations: DP, differential privacy; HE, homomorphic encryption; SMPC, secure multiparty computation.
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. \(\epsilon\) represents the worst case privacy breach in the absence of any such rare breach. If you assume \(\delta = 0\), you will have a pure (\(\epsilon\))-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 (\(\epsilon, \delta\))-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 \(k\) (\(\epsilon, \delta\))-differentially private algorithms, the composed algorithm is at least \((k\epsilon, k\delta)\)-differentially private. Differential privacy also assures the resistance to post-processing theorem which states passing the output of a (\(\epsilon, \delta\))-differentially private algorithm to any arbitrary randomized algorithm will still uphold the (\(\epsilon, \delta\))-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 (\(\sim\) 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, \(\chi^2\) 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.\(^{85–88}\)
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\(^{1}\) 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.\(^{77,101–103}\) 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-anwering 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 privacy-preserving GANs (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 \(\epsilon\) and \(\delta\), 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...
between privacy and utility demands more attention than it has received\textsuperscript{88,90–92}.

Federated Learning

Federated learning\textsuperscript{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 (\textsuperscript{Fig. 1D}).

Federated learning is an iterative process in which each iteration consists of the following steps\textsuperscript{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:

- \textit{Horizontal (sample-based) federated learning}\textsuperscript{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.
- \textit{Vertical (feature-based) federated learning}\textsuperscript{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).
- \textit{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 and production-oriented open source frameworks. Additionally, there are AI platforms whose goal is to apply federated learning to real-world health care settings. 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. For instance, Sheller et al developed a supervised DNN in a federated way for semantic segmentation of brain gliomas from magnetic resonance imaging scans. Chang et al 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 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. VERTICOX is a vertical federated Cox regression model for survival analysis, which employs the alternating direction method of multiplier (ADMM) and is evaluated on acquired immunodeficiency syndrome (AIDS) and breast cancer survival datasets. Similarly, WebDISCO presents a federated Cox regression model but for horizontally distributed survival data. The grid binary logistic regression (GLORE) and the expectation propagation logistic regression (EXPLORER) 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>
<thead>
<tr>
<th>Authors</th>
<th>Year</th>
<th>Technique</th>
<th>Model</th>
<th>Application</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sheller et al</td>
<td>2018</td>
<td>FL</td>
<td>DNN</td>
<td>Medical image segmentation</td>
</tr>
<tr>
<td>Chang et al</td>
<td>2018</td>
<td>FL</td>
<td>Single weight transfer</td>
<td>Medical image segmentation</td>
</tr>
<tr>
<td>Balachandar et al</td>
<td>2020</td>
<td>FL</td>
<td>Cyclic weight transfer</td>
<td>Medical image segmentation</td>
</tr>
<tr>
<td>Nasirigerdeh et al</td>
<td>2020</td>
<td>FL</td>
<td>Linear regression</td>
<td>Genetic associations</td>
</tr>
<tr>
<td>Wu et al</td>
<td>2012</td>
<td>FL</td>
<td>Logistic regression</td>
<td>Genetic associations</td>
</tr>
<tr>
<td>Wang et al</td>
<td>2013</td>
<td>FL</td>
<td>$X^2$ statistics</td>
<td>Genetic associations</td>
</tr>
<tr>
<td>Li et al</td>
<td>2016</td>
<td>FL</td>
<td>Cox regression</td>
<td>Survival analysis</td>
</tr>
<tr>
<td>Dai et al</td>
<td>2020</td>
<td>FL</td>
<td>Support vector machines</td>
<td>Classifying electronic health records</td>
</tr>
<tr>
<td>Huang et al</td>
<td>2018</td>
<td>FL</td>
<td>Adaptive boosting ensemble</td>
<td>Classifying medical data</td>
</tr>
<tr>
<td>Liu et al</td>
<td>2018</td>
<td>FL</td>
<td>Autonomous deep learning</td>
<td>Classifying medical data</td>
</tr>
<tr>
<td>Chen et al</td>
<td>2019</td>
<td>FL</td>
<td>Transfer learning</td>
<td>Training wearable health care devices</td>
</tr>
<tr>
<td>Li et al</td>
<td>2020</td>
<td>FL + DP</td>
<td>DNN</td>
<td>Medical image segmentation</td>
</tr>
<tr>
<td>Li et al</td>
<td>2019</td>
<td>FL + DP</td>
<td>Domain adoption</td>
<td>Medical image pattern recognition</td>
</tr>
<tr>
<td>Choudhury et al</td>
<td>2019</td>
<td>FL + DP</td>
<td>Neural network</td>
<td>Classifying electronic health records</td>
</tr>
<tr>
<td>Constable et al</td>
<td>2015</td>
<td>FL + SMPC</td>
<td>Statistical analysis (e.g., $X^2$ statistics, ...)</td>
<td>Genetic associations</td>
</tr>
<tr>
<td>Lee et al</td>
<td>2019</td>
<td>FL + HE</td>
<td>Context-specific hashing</td>
<td>Learning patient similarity</td>
</tr>
<tr>
<td>Kim et al</td>
<td>2019</td>
<td>FL + DP + HE</td>
<td>Logistic regression</td>
<td>Classifying medical data</td>
</tr>
</tbody>
</table>

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,\textsuperscript{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\textsuperscript{129} developed a horizontally federated toolset for GWAS, called \textit{sPLINK}, which supports Chi-square test, linear regression, and logistic regression. Notably, federated results from \textit{sPLINK} on distributed datasets are the same as those from aggregated analysis conducted with \textit{PLINK}.\textsuperscript{131} Moreover, they showed that \textit{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).\textsuperscript{132–136} Brisimi et al\textsuperscript{132} presented a federated soft-margin support vector machine (\textit{sSVM}) for distributed electronic health records. Huang et al\textsuperscript{133} introduced LoAdaBoost, a federated adaptive boosting method for learning biomedical data such as intensive care unit data from distinct hospitals\textsuperscript{137} while Liu et al\textsuperscript{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.\textsuperscript{138–146} However, to the best of our knowledge, FedHealth\textsuperscript{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.\textsuperscript{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 (\textit{sPLINK}).\textsuperscript{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\textsuperscript{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.\textsuperscript{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 (\textit{\textasciicircum 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.\textsuperscript{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)\textsuperscript{149–153} or they employ HE (FL + HE)\textsuperscript{154,155} SMPC (FL + SMPC) or both DP and HE (FL + DP + HE)\textsuperscript{103,156,157} to securely share the parameters with the server.\textsuperscript{51,158}

In the genomic and biomedical field, several hybrid approaches have been presented recently. Li et al\textsuperscript{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\textsuperscript{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\textsuperscript{159} developed a federated differential privacy mechanism for gradient-based classification on electronic health records.

There are also some studies that incorporate federated learning with cryptographic techniques. For instance, Constable et al\textsuperscript{52} implemented a privacy-protecting structure for federated statistical analysis such as \(\chi^2\) statistics on GWAS while maintaining privacy using SMPC. In a slightly different approach, Lee et al\textsuperscript{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\textsuperscript{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) 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, 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). 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.](image-url)
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 $\chi^2$ 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.\textsuperscript{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.

**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 Schwalbe, Christof Tschoh, and Andrea Wohner for their helpful comments toward improving the paper.

**References**

prediction of sepsis in the ICU. Crit Care Med 2018;46(04):547–553
33 Zerhouni EA, Nabel EG. Protecting aggregate genomic data. Science 2008;322(5898):44–44
46 Jagadeesh KA, Wu DJ, Birgmeier JA, Boneh D, Bejerano G. Keeping patient phenotypes and genotypes private while seeking disease diagnoses. bioRxiv 2019:746230
55 Hasan Z, Mahdi MSR, Mohammed N. Secure count query on encrypted genomic data: a survey. IEEE Internet Comput 2018;22(02):71–82


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


137 Pollard TJ, Johnson AEW, Raffe 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


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