ANNUAL REVIEWS

Annual Review of Biomedical Data Science

Privacy-Enhancing Technologies in Biomedical Data Science

Hyunghoon Cho,[1] David Froelicher,[2,*]
Natnatee Dokmai,[1,*] Anupama Nandi,[1,*]
Shuvom Sadhuka,[2,*] Matthew M. Hong,[2,*]
and Bonnie Berger[2,3]

[1] Department of Biomedical Informatics and Data Science, Yale School of Medicine, New Haven, Connecticut, USA; email: hoon.cho@yale.edu

[2] Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA; email: bab@mit.edu

[3] Department of Mathematics, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA

Annu. Rev. Biomed. Data Sci. 2024. 7:317–43

The *Annual Review of Biomedical Data Science* is online at biodatasci.annualreviews.org

https://doi.org/10.1146/annurev-biodatasci-120423-120107

*These authors contributed equally to this article

Keywords

biomedical data privacy, genomic privacy, privacy-enhancing technologies, secure computation, data sharing, collaborative studies

Abstract

The rapidly growing scale and variety of biomedical data repositories raise important privacy concerns. Conventional frameworks for collecting and sharing human subject data offer limited privacy protection, often necessitating the creation of data silos. Privacy-enhancing technologies (PETs) promise to safeguard these data and broaden their usage by providing means to share and analyze sensitive data while protecting privacy. Here, we review prominent PETs and illustrate their role in advancing biomedicine. We describe key use cases of PETs and their latest technical advances and highlight recent applications of PETs in a range of biomedical domains. We conclude by discussing outstanding challenges and social considerations that need to be addressed to facilitate a broader adoption of PETs in biomedical data science.

1. INTRODUCTION

Data sharing is a vital force in biomedical innovation. Public data repositories and biobanks allow researchers at various organizations to analyze vast arrays of human subject data beyond what they may be able to collect themselves. Many academic labs, commercial enterprises, and hospitals have joined to form collaborative consortia to share biomedical data in the hope of extracting insights that are inaccessible to individual entities due to limited dataset sizes. Policies and guidelines that promote the public dissemination of research data established by government entities (e.g., National Institutes of Health Data Management and Sharing Policy; **https://sharing.nih.gov**) and international standard-setting organizations such as the Global Alliance for Genomics and Health (1) have played pivotal roles in preserving the culture of data sharing among the biomedical community, a tradition rooted in landmark collaborative efforts such as the Human Genome Project.

As we enter the era of personalized medicine, broader sharing of biomedical data is becoming more essential than ever. The limited diversity of human populations represented by existing biomedical datasets has reinforced inequities in how different groups benefit from biomedical advances (2). Studying rare diseases often requires merging small patient cohorts across organizations to enhance statistical power (3). Furthermore, accurately inferring health-related insights for each unique individual requires access to computational models trained on large and multimodal datasets capturing the wide spectrum of individual variation in health and disease. Although recently created biobanks [e.g., the All of Us Research Program (4)] have taken a significant step toward recruiting diverse study participants, these resources are increasingly stored within siloed computing environments, limiting the scope and use of these datasets.

To further expand data-sharing efforts in biomedicine, growing concerns about the privacy risks must be addressed with robust mitigation measures. In the absence of such measures, we may witness a greater dependence on restricted data silos, further compounded by the recent surge in more stringent privacy regulations [e.g., the General Data Protection Regulation (GDPR) in the European Union] and the escalating risks associated with the increasing data scale and computational advances in biomedicine. Moreover, a major data breach has the potential to erode public trust in the scientific enterprise. The loss of trust could not only hinder the efforts to gather large datasets but also worsen inequities by disproportionately affecting the willingness of certain populations to participate in studies.

Privacy-enhancing technologies (PETs) offer promising technical solutions to overcome these challenges by employing a variety of mathematical, algorithmic, and hardware design approaches to enable the sharing and analysis of sensitive data while protecting privacy (**Figure 1**). PETs encompass a broad range of techniques that address different data-sharing scenarios and introduce various trade-offs in terms of the type of supported analyses, computational cost, and the degree of privacy protection offered. In this review, we focus on technologies that are most widely studied in the literature, including homomorphic encryption (HE), secure multiparty computation (MPC), trusted execution environment (TEE), differential privacy (DP), and federated learning (FL). Recent advances have greatly increased the applicability of each of these technologies in biomedicine, as we illustrate in this review. Unlike existing reviews that describe PETs as a potential solution to data-sharing challenges in biomedicine (5–9), we focus on providing an accessible summary of the latest advances in PETs, examining both their technical foundations and biomedical applications.

The rest of this review is organized as follows. Section 2 presents background on biomedical data privacy, covering its historical context and key concepts. Section 3 outlines the scenarios in biomedical research that involve data sharing. Section 4 delves into each PET, providing an overview of techniques, recent advances, limitations, and recent publications, exploring its

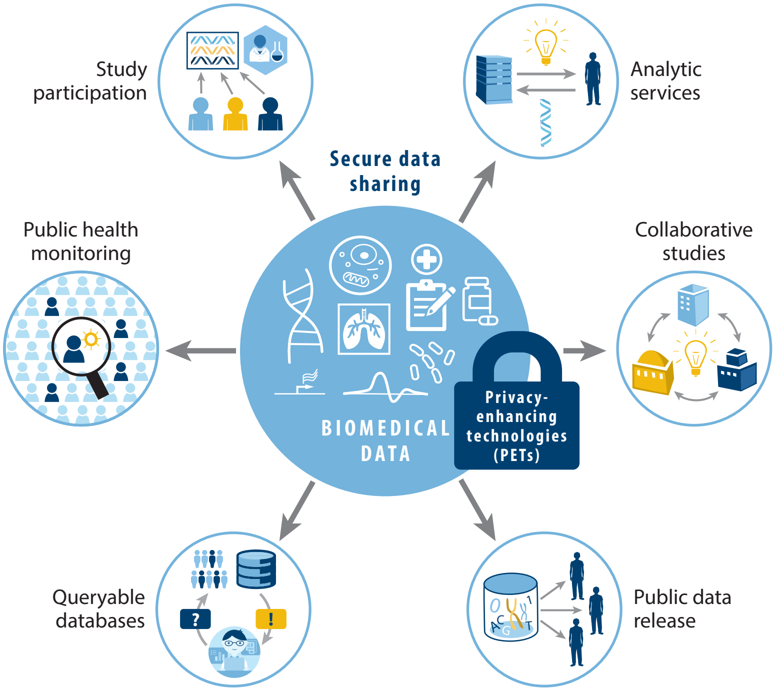

Figure 1

Privacy-enhancing technologies (PETs) provide a range of mathematical, algorithmic, and hardware-based solutions to enable the sharing and analysis of sensitive biomedical data in various settings while protecting data privacy.

application to biomedical tasks. Section 5 reviews related techniques that help facilitate the sharing of biomedical data. Finally, in Section 6, we conclude by discussing open challenges and highlighting key directions for future work.

2. BIOMEDICAL DATA PRIVACY: CHALLENGES AND EXISTING SAFEGUARDS

Data privacy challenges in biomedicine have continually evolved over the past several decades, shaped by technological advances, increasing public awareness, and changes in policies and laws. From the 1960s to the 1980s, the biomedical community saw the establishment of ethical principles in human subject research, exemplified by documents such as the Belmont Report (10). At the same time, concerns about patient privacy increased due to the digitization of medical records. The 1990s saw the establishment of the Common Rule and the Health Insurance Portability and Accountability Act (HIPAA) (11), marking initial efforts to create legal frameworks for safeguarding biomedical data in both research and health care. The completion of the Human Genome Project and the rapid growth of genetic research in the 2000s intensified concerns about the protection of human subjects and their genetic privacy. The 2010s witnessed a surge in privacy concerns in broader societal contexts fueled by the rise of social media, major data breaches, and controversy surrounding government surveillance (12, 13). The international community responded to these concerns by strengthening oversight over the collection, sharing, and use of personal information, such as via GDPR, enacted in 2018. More recent examples include Arizona's and California's 2021 genetic privacy laws (14), which strengthened privacy requirements for storing and sharing genetic data, as well as the National Institute of Standards and Technology's genomic cybersecurity initiative (**https://www.nccoe.nist.gov/projects/cybersecurity-genomic-data**), which called for new security standards in genomics. Currently, privacy concerns persist and extend to new biomedical domains, such as digital health [e.g., electronic health records (EHRs), mobile applications, and wearable devices], multiomics, and epidemic responsiveness in the wake of the COVID-19 pandemic (15). The growing number of large-scale biobanks and data repositories further amplifies these challenges.

The risks posed by biomedical data breaches are multifaceted. Although individuals have different notions of an acceptable privacy risk, unauthorized exposure of private information related to one's biology and health may result in emotional distress, stigmatization, and discrimination in employment, education, and insurance opportunities. Perhaps more concerning, these harms could extend to families and demographic groups, for example, by disclosing genetic relationships between individuals or elevated health risks within groups, respectively. From the perspective of organizations that manage sensitive biomedical data, data breaches can lead to financial penalties, legal consequences, operational disruptions, and reputation damage. These risks are not just hypothetical—as of this writing, a lawsuit has been filed against 23andMe for a data breach that compromised nearly 1 million customers, including their full names, birthdates, and DNA profiles, which were being sold on the dark web for up to $10 per leaked individual (16). Such leaks, if repeated, would lead to a decline in trust in the biomedical research enterprise and health systems, which would further set back future research efforts and impede scientific progress.

Biomedical privacy breaches can occur through multiple routes, each posing unique challenges (8, 17). Phenotype inference aims to deduce an individual's traits or health conditions from different types of biomedical data, often in an unexpected manner. Reidentification occurs when inadequately anonymized data are linked back to an individual. This may involve combining multiple datasets, utilizing auxiliary information, or exploiting vulnerabilities in the deidentification process. Data linkage integrates information from different datasets to construct a more comprehensive profile of an individual, allowing attackers to reveal identity, health status, or other sensitive information about the individual. Even when the complete dataset, including individual-level information, is not directly shared, privacy may still be breached: Data reconstruction attacks attempt to assemble fragments of available information to reveal a portion of the original data; and membership inference attacks focus on determining whether an individual is part of a specific dataset, which can disclose whether the person belongs to a sensitive or stigmatized group, potentially leading to privacy violations.

Conventional approaches for protecting the privacy of biomedical data include policies and laws, technical security measures, and contractual agreements. Legal and financial penalties help serve as deterrents to the misuse of biomedical information; both HIPAA and GDPR prescribe such penalties for noncompliance. The standard practice for securing biomedical data involves encrypting data at rest, employing a secure computing infrastructure, and deidentification strategies (18). Access control and user authentication mechanisms also are widely used to ensure that only authorized individuals can access sensitive data. Furthermore, researchers and organizations that share access to sensitive data typically establish data use agreements (DUAs) to define the permitted use of the data and guidelines for data management. Similarly, business associate agreements, which are contracts between HIPAA-covered entities (e.g., hospitals) and their business associates (e.g., third-party service providers or collaborators), are an important tool for ensuring that entities that handle protected health information comply with data protection standards.

Although these approaches offer useful safeguards for biomedical data, they do not eliminate the possibility of data breaches or reidentification and rely primarily on limiting access to data to achieve security, which has resulted in many datasets being isolated and placed beyond the reach of most researchers. In addition, the ability to detect breaches and enforce penalties can be limited in practice; legal and policy criteria standards that determine what data qualify as private or sufficiently deidentified for sharing lack clear definitions. As a result, many existing datasets are either shared insecurely based on trust or deemed ineligible for sharing due to privacy risks.

3. DATA-SHARING SCENARIOS AND LIMITATIONS

Privacy risks and data-sharing constraints vary across different scenarios in biomedical research and practice. Here, we outline the typical data-sharing scenarios (as illustrated in **Figure 1**) as well as the challenges faced by stakeholders in each context.

3.1. Study Participation

A private individual may voluntarily contribute health data and other personal information to studies (either clinical or nonclinical), data repositories, and third-party services. Participants typically provide informed consent, which outlines the details of data sharing, including the purpose, scope, and potential risks involved. Data sharing benefits participants by improving the collective understanding of health and disease, which could lead to better treatments or other health-related decisions. In the context of some services, participants may also receive personalized data insights. Privacy concerns are a key factor in an individual's decision to participate in a study (19). The adequacy of informed consent remains a topic of ongoing debate, particularly with regard to the ethics of obtaining broad consent for the secondary use of data (20). Effectively communicating privacy risks can also be challenging, as some risks require technical expertise to fully understand them or are poorly understood even among researchers.

3.2. Queryable Databases

Queryable databases are specifically designed to allow users to retrieve information through queries, providing a structured and efficient way to access biomedical data. Examples include patient registries used for study recruitment and various databases containing annotated genetic variants (21), EHRs (22), and clinical trial data (23). While the restricted nature of queries minimizes information leakage, studies have shown that even with these limitations, unintended disclosures can occur (24). Such concerns may discourage individuals from participating in these databases as well as pose data management challenges to the database provider.

3.3. Analytic Services

This refers to the practice of delegating computational tasks to external entities that have access to restricted models or data resources, or more computing resources. The increasing computational demands of biomedical analysis workflows are compelling researchers to increasingly use these third-party services (25), many of which are hosted in cloud environments. However, privacy concerns or regulations can limit the use of these services (e.g., a European Union researcher wishing to upload data to a service operating in a non–European Union country). Moreover, service providers may be required to introduce measures to protect the auxiliary models and data used by the server, which could be leaked through the analysis results returned to users (26).

3.4. Collaborative Studies

There is an increasing need for researchers from different institutions or countries to collaborate on shared research goals by combining their data to obtain a more complete understanding of the biomedical phenomenon of interest. This scenario often involves the establishment of consortia focused on specific health conditions or research areas. However, organizations may need to comply with policies that limit or prohibit external sharing of data, a problem that is exacerbated if entities operate in different regulatory environments or countries.

3.5. Public Data Release

This involves making biomedical datasets and analysis results openly accessible to the broader scientific community and the public. This practice supports the transparency and reproducibility of scientific studies and promotes collaborative efforts such as data science competitions (e.g., Kaggle; **https://www.kaggle.com/**). It also extends the utility of collected datasets by allowing researchers around the world to reanalyze existing data. However, releasing a dataset containing individual-level information poses great privacy risk and is feasible only in rare circumstances. Care must also be taken when releasing simulated or redacted datasets based partially on private data, as these may still lead to privacy leakage.

3.6. Public Health Monitoring

The COVID-19 pandemic motivated the design of public health systems capable of monitoring disease outbreaks and facilitating responses (e.g., exposure notification apps). The effective operation of these systems may require the collection of a broad range of personal information beyond health status, including demographic, geolocation, and social activity data. In addition, sharing these data across jurisdictions may be essential for a more accurate understanding of infectious agents. However, the possibility of harming individuals due to the disclosure of private information remains a significant concern (15), which can prevent the widespread adoption of these systems.

4. PRIVACY-ENHANCING TECHNOLOGIES AND THEIR BIOMEDICAL APPLICATIONS

PETs represent a collection of computational techniques to safeguard sensitive biomedical data. Collectively, these technologies enable the development of privacy-by-design methods for sharing and analyzing biomedical data. These improve both the privacy and utility of biomedical data beyond what is feasible given existing security practices and contractual safeguards, such as DUAs. We describe each technology in detail and discuss recent methodological advances and applications within the biomedical domain.

4.1. Secure Multiparty Computation

MPC allows multiple parties to work together to perform computations collectively on their private inputs without revealing the input to each other. There are two main techniques used for MPC: garbled circuits and secret sharing.

First introduced by Yao in 1986 (27), a garbled circuit enables secure evaluation of a function, represented as a Boolean circuit, between two parties with private input. The input and output of each logical gate in a garbled circuit are randomly masked to prevent the evaluator from gaining information during the circuit evaluation. The input of one party is securely communicated to the other party using the cryptographic primitive of oblivious transfer, which then allows the receiver to evaluate the circuit without knowing the raw input. Although the exponential scaling of the circuit size for complex analysis tasks often leads to high communication and computational costs, several enhancements (28–31) have improved the efficiency of these schemes (32–34).

Secret sharing schemes (35, 36) allow a group of parties to collectively encode a private number by dividing it into random shares, which are held individually by the parties. The private number can be reconstructed only when a predefined number of shares are combined. For example, in additive secret sharing schemes, which are most commonly used in practical settings, secret shares are random elements of a ring (an algebraic structure exemplified by a set of integers modulo a certain number) that add up to the private value. This ensures that all parties' shares must be combined to reveal the secret; any subset reveals no information that can be used to infer the secret. Securely adding two secret-shared numbers, x and y, involves each party adding their individual shares for x and y, resulting in new shares representing $x + y$. Secure multiplication requires interaction between parties (37) but preserves the confidentiality of the private input by masking the numbers shared between parties. Other operations, such as division, square root, and comparison, are performed using addition, multiplication, and special routines that exploit the bitwise representation of private values. These operations can be combined to securely perform various analyses on private data held by multiple parties.

Many frameworks and compilers have been developed to ease the implementation of MPC algorithms leveraging various building block protocols (38). Hybrid schemes (39, 40) and compilers (32, 38, 41), which combine different MPC methodologies to improve efficiency, have also been proposed. For example, ABY (41) proposed switching between garbled circuits and different types of secret sharing (integer or Boolean) to perform each operation in the domain where it is most efficient (e.g., evaluating comparisons in a two-party setting using garbled circuits or evaluating multiplexors or other bitwise operations with more than two parties using Boolean secret sharing). These frameworks have been extended and optimized for applications in machine learning (ML), such as training and inference of neural network models (42–45). Recent enhancements of core operations such as secure comparison (46) have further improved the performance and versatility of MPC frameworks.

The primary limitation of MPC is its substantial communication cost. While garbled circuits allow most of the computation to be performed noninteractively by transferring the entire circuit in a single round of communication, the circuit is typically limited to Boolean operations, and the size of the circuit can become impractically large for sophisticated numerical calculations. Secret sharing enjoys greater analytic flexibility and efficiency in general compared to garbled circuits, but secret sharing–based MPC typically requires many rounds of interaction for complex tasks, a potential bottleneck in limited communication settings (e.g., a wide-area network with large round-trip delays). Furthermore, the requirement that the entire input dataset be secret-shared among the parties can be a hurdle for large-scale biomedical datasets.

Several recent works have developed MPC protocols for a range of analysis tasks in biomedicine (47–52). A common goal of these works is to improve the efficiency of MPC by redesigning the analysis task in a way that is more amenable to efficient computation using MPC operations. For example, Cho et al. (47) introduced a generalization of secret sharing techniques aimed at minimizing redundant computation, which led to an efficient algorithm for genome-wide association studies (GWAS), involving sophisticated linear algebra tasks such as principal component analysis. This work was extended to address collaborative prediction of drug-target interactions using a neural network model (53). Jagadeesh et al. (49) used garbled circuits to efficiently perform Boolean operations (such as set intersection and difference) to identify genetic variants of interest in patient genomes. Von Maltitz et al. (54) introduced an MPC protocol for survival analysis based on the Kaplan-Meier estimator. A different approach was taken by Smajlović et al. (55), who developed a Python-based compiler that transforms a high-level analysis code into MPC executables incorporating automated optimization based on static code analysis. Such tools can help accelerate the development of MPC applications for various biomedical tasks by making the techniques more accessible to biomedical practitioners.

4.2. Homomorphic Encryption

HE refers to a form of encryption that allows direct computations on encrypted data. Earlier HE schemes, such as those by Rivest et al. (56), Elgamal (57), Paillier (58), and Goldwasser & Micali (59), were known as leveled or somewhat HE schemes, supporting specific types of operations, for example, additions only or multiplications only, or a limited number of them. In 2009, Gentry (60) introduced the first construction of a fully homomorphic encryption (FHE) scheme that allows arbitrary arithmetic computations through a bootstrapping technique, which refreshes a ciphertext (encrypted data) to support additional operations. To address the limited concrete efficiency of the initial scheme of Gentry, which required several minutes of runtime for each bit operation (61), several schemes were later proposed (62–65) that reduced the overall computational cost of FHE and thereby enabled its use in practical applications.

Akin to standard encryption schemes, the security of HE is based on the hardness of well-studied mathematical problems. Many HE schemes are based on the ring learning with errors (RLWE) problem (66, 67), a lattice-based problem where the goal is to distinguish whether a set of ring elements is sampled randomly or approximately the result of multiplying a known set of elements with a common secret element. This problem is shown to be extremely difficult to solve without knowing the secret but is otherwise easy, translating into the guarantee that an entity can decrypt a ciphertext only if the decryption key is known. The random noise that is introduced into the ciphertext to maintain the difficulty of this problem increases with each homomorphic operation. Unlike schemes that precisely perform computation at the expense of reducing the range of encoded values (62–64), the CKKS scheme by Cheon et al. (65) adds noise directly to the data values, enabling efficient operations at a small loss in precision. CKKS has been widely adopted in scientific applications where a small amount of noise can be tolerated. In all RLWE-based schemes, a single ciphertext encodes multiple values, and homomorphic operations such as addition and multiplication are performed simultaneously on all values in a ciphertext—known as the single-instruction, multiple-data property. Exploiting this property can improve the scalability of these schemes.

Notable recent developments include more efficient bootstrapping techniques (68, 69) and alternative constructions that offer a trade-off between different types of operations. For example, the TFHE scheme by Chillotti et al. (70) is constructed based on the mathematical structure of the torus and permits efficient bootstrapping yet is limited to Boolean or bitwise operations. Furthermore, several HE compilers have been proposed (71) to streamline the development and optimization of HE algorithms, for example, to simplify the management of ciphertext noise. Tailored frameworks have also been developed for the secure training of predictive ML models (72, 73).

In the biomedical domain, HE has mainly found applications in the outsourcing of computational tasks involving sensitive data. These computations may be challenging for individual users to perform because of the scale of the problem (in terms of both dataset size and computational complexity) or because of limited access to additional data or models required for the analysis. HE helps to ensure that the user's data remain private when analysis is delegated to a third party. For example, HE-based solutions have been proposed for privately outsourcing the detection of heart conditions in electrocardiogram data (74), as well as cardiovascular risk prediction based on health records (75). Many works have tackled the computation of GWAS statistics on encrypted data, addressing a range of statistics and application settings (76–80). Other tasks explored in the literature

EXTENDING HOMOMORPHIC ENCRYPTION TO COLLABORATIVE ANALYSIS SETTINGS: MULTIPARTY HOMOMORPHIC ENCRYPTION

A recent line of work explores a novel use of HE to facilitate collaborative studies. Conventional HE schemes, described in Section 4.2, allow any party with the decryption key to access the private data. In contrast, threshold HE schemes (87–90) use a decryption key that is secret-shared among a group of parties, allowing them to individually operate over encrypted data while ensuring that only the data values that are agreed upon among the parties can be decrypted. Similarly, multi-key schemes (91, 92) allow each party to use their own key and modify operations to support data encrypted under different keys.

These *multiparty* HE (MHE) schemes open the door to HE-based algorithms that can analyze private data distributed among multiple parties, analogously to the MPC setting. Recent studies (93, 94) have shown that these schemes can enable seamless integration of HE operations with efficient interactive routines, including MPC protocols, to reduce the cost of challenging operations such as bootstrapping (89). Importantly, these schemes allow each party to leverage efficient local computations using the locally available unencrypted data. MHE can thus help address the scalability limitations of standalone applications of HE or MPC by offering a federated analysis paradigm in which it is necessary neither to secret-share the entire private dataset among the parties nor to encrypt and centralize all data at a single site for analysis.

Applications of MHE are being explored in various domains, including distributed ML and linear algebra (95–98) as well as collaborative biomedical analyses, such as GWAS (88, 93, 94, 99) and cell type classification (100). Recent results (94) demonstrate the practicality of this approach in handling complex biomedical tasks on the scale of modern biobanks that include hundreds of thousands of individuals. However, addressing each application currently requires time-consuming effort to design and optimize algorithms to achieve practical runtimes. Ongoing work on streamlining the development and use of these solutions, e.g., through cloud-based analysis platforms (101) and easy-to-use programming frameworks or libraries (55, 102), can help make these tools more widely available.

include count queries on genomic and medical databases (e.g., for cohort exploration) (81, 82), detection of genetic parent-child relationships (83), and disease risk prediction using both clinical and genomic information (74, 84). Finally, Kim et al. (85) and Gürsoy et al. (86) recently illustrated secure imputation of an encrypted private genome.

These advances have brought HE-based solutions closer to meeting the requirements of biomedical applications. Nevertheless, the scope of these applications remains restricted due to several factors, including the substantial computational overhead of homomorphic operations compared to unencrypted analysis, the need to approximate nonlinear operations using additions and multiplications, and the practical limits on the complexity of the analysis task due to the high cost of bootstrapping. Moreover, most of the aforementioned solutions require that all input data be encrypted and transferred to the entity performing the computation, which can be a significant burden for large datasets. In the sidebar titled Extending Homomorphic Encryption to Collaborative Analysis Settings: Multiparty Homomorphic Encryption, we describe a recent technical advance that helps address these limitations.

4.3. Trusted Execution Environments

A TEE is a secure area within the main processor, also called an enclave, that ensures the safe and isolated execution of software. This isolation guarantees that the memory content, end-to-end communication with external parties, and control flow of the application are protected from untrusted or malicious processes running within the same hardware, including a malicious operating system or hypervisor (103, 104). In certain TEE architectures, the binary executable of an application can also be verified through a process called remote attestation (105). To achieve these security properties, TEE relies on core hardware security components built into the processor that cannot be manipulated by software. These components typically comprise a memory encryption engine and controller to isolate memory access and integrated circuits for cryptographic key storage and operation.

Recent TEE developments have focused on supporting third-party software deployment in an untrusted cloud environment, addressing the deployment of both user-level applications and virtual machines (VMs). Popular TEE platforms include Intel Software Guard Extension (SGX) (106) for user-level applications and Intel Trust Domain Extensions (107) and AMD Secure Encrypted Virtualization (108) for VMs. Nvidia recently introduced an update to its graphics processing unit (GPU) architecture that enables GPU computation in TEE (109). In a mobile setting, Arm TrustZone (104) is a ubiquitous TEE platform on Arm central processing units (CPUs), but generally a limited set of TEE functionalities are available for mobile applications.

Although TEEs offer the capability to confidentially analyze private data with computational efficiency and functionalities similar to conventional computing environments, their major drawback lies in the complexities of achieving hardware-based security. Unlike MPC and HE, which rely on minimal and well-established cryptographic primitives, TEEs' hardware-based approach introduces unique vulnerabilities. Some vulnerabilities are the result of CPU architectural bugs that allow a malicious process to extract protected data from an enclave (110); manufacturers typically promptly patch these problems once they are discovered. Other limitations are inherent in the TEE architecture and lead to the issue of side channels—indirect pathways for information leakage (111, 112). For example, a TEE enclave's access patterns to memory pages can inadvertently reveal sensitive information stored in the secure area to an attacker. Although such attacks require significant effort, software-level mitigation is necessary when the highest level of security is required. One strategy involves ensuring that the memory access or timing patterns of the program do not depend on sensitive information (113). However, such mitigation can incur an additional computational burden and require relevant expertise during algorithm development.

Despite these drawbacks, TEEs have a promising future with vested interests from major CPU producers such as Intel, AMD, and Arm, who continue to address security issues and improve their TEE platforms. Cloud service providers such as Google Cloud Platform and Microsoft Azure offer TEE-enabled computing infrastructure. In addition, initiatives such as the Confidential Computing Consortium (**https://confidentialcomputing.io**) and the Trusted Execution Environment Provisioning (TEEP) Working Group (**https://datatracker.ietf.org/wg/teep/about**) of the Internet Engineering Task Force have been formed to support open-source projects and the development of standards related to TEE. In the research community, many software tools have been introduced in recent years to ease the translation of existing software to run securely on TEE platforms (112).

In the biomedical domain, TEEs have gained significant traction due to their ability to securely outsource the analysis of biomedical data and to facilitate the development and deployment of health artificial intelligence (AI) tools on a large scale. Notable real-world examples include BeeKeeperAI (114), a privacy-preserving health care AI company, and AOK, a network of 11 regional health insurers in Germany. These organizations utilize Intel SGX to protect confidential patient data, complying with regulations such as HIPAA, GDPR, and Germany's Patient Data Protection Act (115). Applications of TEE in genomics are also emerging. An example is a federated GWAS service based on Intel SGX that securely aggregates data from multiple sites and incrementally updates the statistics as study participants are added or removed (116). Data sketching techniques for enhancing the efficiency of GWAS computation in Intel SGX have also been proposed (117). Considering other genome analysis tasks, Widanage et al. (118) demonstrated read mapping in Intel SGX and described a generalization of their tool to other workflows. Dokmai et al. (113) proposed a TEE-based service for secure genotype imputation, introducing techniques to achieve resilience against side channels while maintaining accurate imputation performance.

4.4. Differential Privacy

DP is a mathematical definition of privacy that provides rigorous privacy protection by ensuring that the removal or addition of a single individual in a dataset does not lead to a distinguishable change in the analysis results (119, 120). Formally, given $\varepsilon \geq 0$, a randomized mechanism $\mathcal{A}$ satisfies ε-DP if, for all datasets D_1 and D_2 that differ in one record, and for any subset $\mathcal{O}$ of all possible outputs of $\mathcal{A}$, we have $\mathrm{P}\left[\mathcal{A}(D_1) \in \mathcal{O}\right] \leq e^{\varepsilon}\,\mathrm{P}\left[\mathcal{A}(D_2) \in \mathcal{O}\right]$—intuitively, this means that any result is similar in likelihood between similar datasets. The parameter ε is called the privacy budget and is used to specify the level of privacy protection. DP mechanisms generally satisfy the privacy guarantee by adding noise to the data, where a smaller ε provides more privacy at the cost of greater loss in accuracy by adding more noise. Standard DP techniques include Laplace, Gaussian, and exponential mechanisms, representing different approaches to sampling the noisy analysis result.

Various techniques have been developed to minimize noise addition and obtain a more desirable trade-off between privacy and utility. For example, some DP formulations relax the notion of privacy for better utility: (ε, δ)-DP, also known as approximate DP (120), requires that ε-DP be satisfied with a probability at least $1 - \delta$. Concentrated DP (121), zero-concentrated DP (122), and Rényi DP (123) view privacy loss as a random variable and bound the average loss instead of the worst-case loss.

Key properties of DP include postprocessing, which ensures that further analysis of data that satisfy DP does not result in any additional privacy leakage, and composition, which allows multiple mechanisms operating on the same data to be combined to provide a joint DP guarantee. As a result of these properties, adding DP noise to different components of the analysis pipeline—for example, input, output, optimization objectives (124, 125), or gradients (126, 127)—can have a significant impact on overall precision depending on the analysis task. Another key factor that influences the amount of noise is sensitivity, which measures the maximum change in the analysis output due to a single-record change in the data. Different approaches have been proposed to analyze the sensitivity of a given function [e.g., global, local, or smooth sensitivity (128)]. Due to these considerations, it is often necessary to carefully design DP mechanisms for specific applications to optimize their performance.

For example, in a multiparty setting, DP can be implemented either locally, by individual data providers, or globally, by a central server that aggregates analysis results; these approaches are called local differential privacy (LDP) and centralized differential privacy (CDP), respectively. Although CDP typically requires less noise by adding it directly to aggregated data, it may be more vulnerable to privacy leakage because it relies on a trusted third party for data aggregation. On the other hand, LDP offers DP at the level of individual data providers while increasing the overall amount of noise. Common data perturbation techniques to achieve LDP include the randomized response and its variants (129–131).

DP has recently been deployed by various entities to address private collection of statistics and publication of privatized datasets. The RAPPOR technology by Erlingsson et al. (132) uses randomized response and bloom filters to privately collect usage statistics from the Chrome browser. LDP has been deployed by Apple to collect information about emojis and search queries from its devices (133) and by Microsoft for application-level telemetry in Windows 10 (134). In 2020, the US Census Bureau released the census data with DP using the TopDown algorithm (135), which hierarchically aggregates statistics based on geographic units.

A key focus of DP applications in biomedicine has been on the release of GWAS statistics. Uhlerop et al. (136) introduced DP mechanisms for releasing minor allele frequencies and χ^2 statistics for case-control GWAS. This work was later extended by Yu et al. (137, 138) to handle larger cohorts and logistic regression. An alternative approach based on the exponential mechanism has also been proposed (139). Simmons & Berger (140) introduced an optimization framework for privately reporting a fixed number of the most significant associations. In subsequent work, Simmons et al. (141) developed DP methods for GWAS with correction for population stratification. Other notable applications of DP include the sharing of genotypic data (142), clinical trial data (143), and tabular medical records (144). DP has also been applied in interactive database settings, for example, for count or membership queries (145, 146) and genetic matching of patients (147). In the public health domain, DP has been used to support the development of the COVID-19 Real-Time Information System for Preparedness and Epidemic Response (148) and a mobile diagnostic system for coronary heart disease (149).

Despite these advances, the practical adoption of DP faces several technical challenges. Privacy parameters (e.g., ε) associated with DP methods are an important factor controlling the trade-off between privacy and utility; however, there are no rigorous methods or standards for choosing an acceptable value of these parameters for a given task. Since biomedical data are typically high-dimensional, a large number of statistics need to be shared privately. Moreover, these data are often analyzed using sophisticated algorithms comprised of many steps where DP could be incorporated. As a result, designing effective DP mechanisms that optimally distribute the privacy budget can be difficult. Another limitation is that DP cannot protect every dataset; for example, small datasets typically require an overwhelming amount of noise for DP and need to be protected using other strategies.

4.5. Federated Learning

FL allows multiple parties to collaboratively train ML models in a distributed manner (150). The parties share the model parameters or updates (e.g., gradients) during training but do not directly share the training data, hence mitigating privacy risks. Two main categories of FL use cases include (*a*) cross-silo, in which a small number of parties hold a substantial fraction of the data, and (*b*) cross-device, where a large number of devices (possibly millions) hold a small amount of data (150). The former is more similar to traditional MPC settings where parties may represent different institutions, each of which has collected data from many individuals, while the latter is often found in consumer applications, where, for instance, millions of mobile phones may collect personal user data.

In FL, each party is limited to their local share of the data in evaluating and updating the model; thus, several approaches exist for synchronizing the state of the model across the parties. The federated averaging technique asks each party to locally compute model updates that are sent to a central server to be averaged and applied globally (151). The weights used to average these updates are typically chosen as a function of the size and quality of each party's data (152, 153). More advanced methods such as federated matched averaging synchronize weights layer-wise via matching to cope with permutation invariance in neural networks (154). Other approaches avoid global synchronization and instead iteratively pass weights from party to party (155). Personalized FL is another approach, whereby each party learns a different local model that incorporates both the information from other parties and local data characteristics (156, 157).

The robustness of FL is a major challenge in practice. Issues such as network connectivity, communication constraints, and resource constraints can prevent certain parties from fully participating in every round of the protocol (158, 159). Heterogeneity across data silos or devices may also introduce concerns about inequity and limited generalization of trained models; for instance, simple averaging techniques have been shown to lead to inaccurate results in small subpopulations (160–162).

Another challenge is that FL may provide limited privacy and security protection. For example, by inspecting the model updates from other parties during multiple rounds of the protocol (163), one hospital may be able to infer the characteristics of patients in another hospital. This could reveal information such as the distribution of clinical labels, individual coordinates of feature vectors, and sometimes even entire training inputs (164–166). Furthermore, a malicious adversary could manipulate the data or the model to further their own goals at the expense of others (167).

The recent literature on FL introduces a wide range of techniques to address these limitations. Combining FL with DP can provide rigorous bounds on privacy leakage (168, 169), although doing so while maintaining the accuracy of the model can be challenging. If the central aggregator is not trusted, parties may choose to use LDP to add noise to their local gradients before aggregation (170). Alternatively, MPC, HE, or TEE can also support secure aggregation of model weights so that no additional information is leaked other than aggregated results (163). Solutions that protect model parameters throughout the entire training procedure using encryption techniques have also been proposed (96–98). Robustness is often addressed by adapting the protocol based on the qualities of each party. For example, some methods propose to detect and remove outliers to learn from a core set of reliable parties. Others propose to alter the averaging weights to produce fairer global models that perform comparably well on each party (162). Although existing FL applications focus mainly on supervised learning, recent work extends FL to address other ML tasks, including semisupervised, unsupervised, and reinforcement learning (171–173).

FL has touched upon numerous biomedical applications. In the cross-silo setting, FL can improve analytics and care for patients at various stages of health care by putting together more extensive training data to improve the performance of ML models. Notable uses include rare disease analysis (174, 175), multihospital collaboration for medical image analysis (176–178), and automated phenotyping and risk prediction from clinical notes (179–181). In the cross-device setting, FL has the potential to transform mobile health (182). For example, FL can allow wearable devices, such as Fitbits or Apple Watches, to adapt over time to the individual's unique health and lifestyle characteristics, such as resting heart rate, steps per day, and blood oxygen levels. These models can enable more accurate health monitoring for individuals, for example, for gait identification and fall detection (183, 184).

5. OTHER RELATED TECHNIQUES

Several workflows involving the exchange of private data have received special attention from the privacy and security community to develop targeted methods that extend beyond the scope of PETs described in Section 4. In this section, we highlight some of these techniques.

5.1. Private Information Retrieval

In private information retrieval (PIR), a client retrieves specific items of interest from a database stored on a server without revealing the identity of the accessed items (185, 186). A naïve approach of downloading the entire database and querying it locally is impractical for large datasets. In an HE-based solution, the client uploads an encrypted query, the server searches the database homomorphically, and then it returns the result to the client for decryption. With practical lattice-based HE (Section 4.2) and database preprocessing and amortization techniques (187, 188), recent PIR protocols have been shown to scale to databases that include billions of entries (189–192). Other works have extended PIR to more sophisticated queries such as keyword search in sparse databases (193, 194) and batch querying (195, 196). In the biomedical context, PIR can enhance the utility of public data resources that require users to either download the entire database or disclose private data (e.g., genetic mutations or patient records) to the server in order to query the database. For example, PIR solutions have been proposed for outsourced storage of genomic data, which support secure retrieval of variants of interest (197, 198).

5.2. Private Set Intersection

Private set intersection (PSI) addresses a problem closely related to PIR, where two parties, each holding a set of items, wish to learn the intersection between the two sets without revealing any other information to each other. PSI-size is a notable variant of PSI, where only the size of the intersection is revealed. PSI has been extensively studied, leading to practical protocols for billions of items and several variants addressing different trust assumptions, trade-offs between communication and computation, and number of parties (199–201). Several works have proposed PSI protocols for the computation of genome similarity, viewing each genome as a set of variants: Baldi et al. (202) introduced paternity testing based on PSI techniques (203, 204), and Wang et al. (205) developed a PSI-based protocol for securely calculating the edit distance between genomes.

5.3. Zero-Knowledge Proofs

Verification of computation, an essential component of trustworthy data analytics systems, can be challenging when sensitive biomedical data are involved. A zero-knowledge proof (ZKP) (206) is a cryptographic primitive, related to MPC and digital signatures (56, 207), that allows one to prove the truthfulness of a statement about the data or the computation without disclosing sensitive information. For example, Goldreich et al. (208) showed that generic ZKPs (209) can be used to prove that a secure computation protocol (e.g., MPC) is carried out honestly without decrypting any intermediate value. Although generic constructions typically incur impractical computational overheads, recent advances have improved the efficiency of ZKPs under a variety of security and model assumptions (210–213). Froelicher et al. (214) demonstrated that ZKPs for discrete logarithms (215) can ensure the integrity of certain HE computations in a distributed health analytics system. Chatel et al. (216) introduced a ZKP scheme based on the MPC-in-the-head paradigm (217, 218), allowing direct-to-consumer analytic service providers to verify that the user's uploaded data are from a trusted source, thus preventing a malicious user from tampering with the analysis result.

5.4. Blockchain

Blockchain provides a decentralized framework for securely recording and verifying transactions in a distributed network. It uses cryptographic techniques to create a chain of blocks, that is, a time-stamped list of transactions, providing transparency, immutability, and accountability in data management. Beyond well-known applications in finance (e.g., Bitcoin), blockchain has become increasingly relevant in biomedical domains (219). A key use case is to create a secure and decentralized health information exchange to improve the management of medical records and insurance claims among various stakeholders (220). It can also be used to create a data-sharing platform to support biomedical research while providing data provenance and accountability (221). Privacy protection of data exchanged through blockchains is a key challenge that often requires blockchains to be carefully combined with other encryption techniques or PETs. Another focus of ongoing research is on improving the scalability and robustness of blockchain networks, which is necessary for their deployment across a large network of institutions.

5.5. Synthetic Data Generation

Creating synthetic data that resemble real data without being directly linked to private individuals has become a useful privacy-aware data-sharing strategy (222). Public sharing of synthetic data can support collaborative efforts, such as data analysis competitions and validation of computational models across institutions. It can also support various academic and educational activities, for instance, by creating a realistic patient profile for use in training or public communication. Techniques for generating synthetic data have evolved alongside ML advances, particularly in deep generative models. The introduction of generative adversarial networks and diffusion models greatly improved the synthesis of various types of biomedical data, including medical images (223, 224) and EHRs (225). However, the possibility that synthetic data can leak private information about the original data used to train the models remains a major concern (226). Recent studies have suggested that the greater expressiveness of modern generative models, in fact, increases the likelihood of private training data being reconstructed (227). Although incorporating DP into model training can help mitigate these risks (228), it can degrade the quality of the generated data, especially for high-dimensional data such as images and genomes. Future improvements in both the quality and privacy of synthetic data will be crucial in expanding their use in settings where direct sharing of data is necessary.

6. OPEN CHALLENGES AND OUTLOOK

As the field of biomedical data science expands to encompass a wider variety of data modalities, more complex statistical models, and evolving computing environments, our understanding of privacy risks must also change. Studies that uncover novel privacy risks in emerging data types [e.g., transcriptomics (84, 229–231), proteomics (232), and wearable devices (233)] and computational models [e.g., diffusion and large language models (227, 234)] will be particularly valuable. Integrating these findings into practical guidelines and policies will require a thoughtful examination of the evolving incentives and capabilities of potential adversaries (235–237).

A critical aspect of PETs is the varying degrees of privacy protection they offer and how they can be aligned with our social values and the needs of practitioners. While acknowledging the value of cryptographic PETs (i.e., HE, MPC, and DP) that offer the strongest, formal notions of privacy, we must also be aware of potential pitfalls in the practical implementation of these techniques, such as software flaws (238) or violations of model assumptions (239). Technologies that offer less formal but more widely applicable privacy enhancements (i.e., TEE and FL) can be useful alternatives in some settings. A promising future direction is to explore a joint use of PETs to combine their strengths while mitigating their weaknesses, as described in the sidebar titled Extending Homomorphic Encryption to Collaborative Analysis Settings: Multiparty Homomorphic Encryption. Future policies and regulations will have a crucial role to play in translating the complex privacy properties of emerging tools based on PETs into concrete guidelines for the biomedical community.

The social impact of PETs involves another key consideration: equity (240). Many studies have shown that there are inequities in emerging clinical applications of computational tools (241, 242). Rectifying these issues requires greater data sharing to create more diverse datasets, which in turn introduces new privacy challenges (243, 244). On the other hand, those whose data are most needed to improve equity in biomedicine (e.g., underrepresented groups) may also have the most to lose in the event of a privacy breach (245). Furthermore, certain PETs, such as DP, may disproportionately reduce the accuracy of ML models in populations with limited representation in the dataset (246). Navigating this complex trade-off between privacy and equity remains an important challenge.

We expect trust and transparency to play a crucial role in aligning PETs with the interests of stakeholders in organizational settings (247). PETs can be viewed as a tool to strengthen trust between stakeholders by increasing transparency and mitigating various privacy and security risks that emerge in collaborative partnerships. This perspective contrasts with a common focus in the PETs community on preventing malicious actors from breaching systems and gaining access to sensitive data. Integrating contextual values such as trust and human-centered design principles into PETs could foster the creation of tools that more effectively address the needs of the biomedical community.

As PETs continue to mature and become more broadly applicable, as demonstrated in this review, there will be a growing need to tailor these techniques to create effective algorithms and tools that address diverse biomedical workflows. A closer collaboration among PET developers, biomedical practitioners, policymakers, and patients and study participants could help prioritize efforts that address the most pressing challenges. Furthermore, software development and deployment tools designed to assist researchers in incorporating PETs into their existing workflows could help ensure that these techniques are broadly accessible. The combination of advances in foundational techniques, effective algorithm design, and the establishment of social frameworks to safeguard the use of these technologies will be key to unlocking the potential of PETs in biomedical data science.

DISCLOSURE STATEMENT

B.B. and H.C. have pending patent applications US20230154630A1 and US20210398611A1. The authors declare no other potential conflicts of interest.

ACKNOWLEDGMENTS

This work was supported by National Institutes of Health grants R01 HG010959 (to B.B.) and RM1 HG011558 and DP5 OD029574 (to H.C.) and by National Science Foundation Graduate Research Fellowship grant 2141064 and the Hertz Foundation Fellowship (to S.S.). We thank Simon Mendelsohn for comments.

LITERATURE CITED

1. Rehm HL, Page AJH, Smith L, Adams JB, Alterovitz G, et al. 2021. GA4GH: international policies and standards for data sharing across genomic research and healthcare. *Cell Genom.* 1(2):100029
2. Fatumo S, Chikowore T, Choudhury A, Ayub M, Martin AR, Kuchenbäcker K. 2022. Diversity in genomic studies: a roadmap to address the imbalance. *Nat. Med.* 28(2):243–50
3. Philippakis AA, Azzariti DR, Beltran S, Brookes AJ, Brownstein CA, et al. 2015. The matchmaker exchange: a platform for rare disease gene discovery. *Hum. Mutat.* 36(10):915–21
4. All of Us Res. Program Investig. 2019. The "All of Us" research program. *N. Engl. J. Med.* 381(7):668–76
5. Arellano AM, Dai W, Wang S, Jiang X, Ohno-Machado L. 2018. Privacy policy and technology in biomedical data science. *Annu. Rev. Biomed. Data Sci.* 1:115–29
6. Gürsoy G. 2022. Genome privacy and trust. *Annu. Rev. Biomed. Data Sci.* 5:163–81
7. Wan Z, Hazel JW, Clayton EW, Vorobeychik Y, Kantarcioglu M, Malin BA. 2022. Sociotechnical safeguards for genomic data privacy. *Nat. Rev. Genet.* 23(7):429–45
8. Bonomi L, Huang Y, Ohno-Machado L. 2020. Privacy challenges and research opportunities for genomic data sharing. *Nat. Genet.* 52(7):646–54
9. Berger B, Cho H. 2019. Emerging technologies towards enhancing privacy in genomic data sharing. *Genome Biol.* 20(1):128

10. Beauchamp TL. 2008. The Belmont Report. In *The Oxford Textbook of Clinical Research Ethics*, ed. EJ Emanuel, C Grady, RA Crouch, RK Lie, FG Miller, D Wendler, pp. 149–55. New York: Oxford Univ. Press
11. Nosowsky R, Giordano TJ. 2006. The Health Insurance Portability and Accountability Act of 1996 (HIPAA) privacy rule: implications for clinical research. *Annu. Rev. Med.* 57:575–90
12. Isaak J, Hanna MJ. 2018. User data privacy: Facebook, Cambridge Analytica, and privacy protection. *Computer* 51(8):56–59
13. Greenwald G. 2014. *No Place to Hide: Edward Snowden, the NSA, and the US Surveillance State*. New York: Metropolitan Books
14. McKeon J. 2021. Growing number of states enact new genetic data privacy laws. *TechTarget*, Oct. 27. **https://healthitsecurity.com/news/growing-number-of-states-enact-new-genetic-data-privacy-laws**
15. Cho H, Ippolito D, Yu YW. 2020. Contact tracing mobile apps for COVID-19: privacy considerations and related trade-offs. arXiv:2003.11511 [cs.CR]
16. Adler S. 2023. First lawsuit filed over 23andMe data breach. *HIPAA Journal*, Oct. 12. **https://www.hipaajournal.com/first-lawsuit-filed-over-23andme-data-breach/**
17. Erlich Y, Narayanan A. 2014. Routes for breaching and protecting genetic privacy. *Nat. Rev. Genet.* 15(6):409–21
18. Garfinkel S. 2015. *De-identification of personal information*. Rep. 8053, Natl. Inst. Stand. Technol., Gaithersburg, MD. **https://nvlpubs.nist.gov/nistpubs/ir/2015/nist.ir.8053.pdf**
19. Clayton EW, Halverson CM, Sathe NA, Malin BA. 2018. A systematic literature review of individuals' perspectives on privacy and genetic information in the United States. *PLOS ONE* 13(10):e0204417
20. Steinsbekk KS, Kåre Myskja B, Solberg B. 2013. Broad consent versus dynamic consent in biobank research: Is passive participation an ethical problem? *Eur. J. Hum. Genet.* 21(9):897–902
21. Fiume M, Cupak M, Keenan S, Rambla J, de la Torre S, et al. 2019. Federated discovery and sharing of genomic data using beacons. *Nat. Biotechnol.* 37(3):220–24
22. Fleurence RL, Curtis LH, Califf RM, Platt R, Selby JV, Brown JS. 2014. Launching PCORnet, a national patient-centered clinical research network. *J. Am. Med. Inform. Assoc.* 21(4):578–82
23. Zarin DA, Tse T, Williams RJ, Califf RM, Ide NC. 2011. The ClinicalTrials.gov results database—update and key issues. *N. Engl. J. Med.* 364(9):852–60
24. Shringarpure SS, Bustamante CD. 2015. Privacy risks from genomic data-sharing beacons. *Am. J. Hum. Genet.* 97(5):631–46
25. Das S, Forer L, Schönherr S, Sidore C, Locke AE, et al. 2016. Next-generation genotype imputation service and methods. *Nat. Genet.* 48(10):1284–87
26. Mosca MJ, Cho H. 2023. Reconstruction of private genomes through reference-based genotype imputation. *Genome Biol.* 24(1):271
27. Yao AC-C. 1986. How to generate and exchange secrets. In *27th Annual Symposium on Foundations of Computer Science (SFCS 1986)*, pp. 162–67. Piscataway, NJ: IEEE
28. Malkhi D, Nisan N, Pinkas B, Sella Y. 2004. *Fairplay—a secure two-party computation system*. Paper presented at USENIX Security Symposium, San Diego, CA, Aug. 9
29. Kolesnikov V, Schneider T. 2008. Improved garbled circuit: free XOR gates and applications. In *Automata, Languages and Programming: 35th International Colloquium, ICALP 2008, Reykjavik, Iceland, July 7–11, 2008, Proceedings, Part II*, ed. L Aceto, I Damgård, LA Goldberg, MM Halldórsson, A Ingólfsdóttir, I Walukiewicz, pp. 486–98. Berlin: Springer
30. Pinkas B, Schneider T, Smart NP, Williams SC. 2009. Secure two-party computation is practical. In *Advances in Cryptology—ASIACRYPT 2009: 15th International Conference on the Theory and Application of Cryptology and Information Security, Tokyo, Japan, December 6–10, 2009, Proceedings*, ed. M Matsui, pp. 250–67. Berlin: Springer
31. Huang Y, Evans D, Katz J, Malka L. 2011. *Faster secure two-party computation using garbled circuits*. Paper presented at the 20th USENIX Security Symposium, San Francisco, CA, Aug. 10
32. Songhori EM, Hussain SU, Sadeghi AR, Schneider T, Koushanfar F. 2015. Tinygarble: highly compressed and scalable sequential garbled circuits. In *2015 IEEE Symposium on Security and Privacy*, pp. 411–28. Piscataway, NJ: IEEE

33. Liu C, Wang XS, Nayak K, Huang Y, Shi E. 2015. Oblivm: a programming framework for secure computation. In *2015 IEEE Symposium on Security and Privacy*, pp. 359–76. Piscataway, NJ: IEEE

34. Rastogi A, Hammer MA, Hicks M. 2014. Wysteria: a programming language for generic, mixed-mode multiparty computations. In *2014 IEEE Symposium on Security and Privacy*, pp. 655–70. Piscataway, NJ: IEEE

35. Shamir A. 1979. How to share a secret. *Commun. ACM* 22(11):612–13

36. Blakley GR. 1979. Safeguarding cryptographic keys. In *International Workshop on Managing Requirements Knowledge*, pp. 313–18. Piscataway, NJ: IEEE Comp. Soc.

37. Beaver D. 1992. Efficient multiparty protocols using circuit randomization. In *Advances in Cryptology—CRYPTO '91*, ed. J Feigenbaum, pp. 420–32. Berlin: Springer

38. Hastings M, Hemenway B, Noble D, Zdancewic S. 2019. Sok: general purpose compilers for secure multi-party computation. In *2019 IEEE Symposium on Security and Privacy (S&P)*, pp. 1220–37. Piscataway, NJ: IEEE

39. Keller M. 2020. MP-SPDZ: a versatile framework for multi-party computation. In *CCS '20: Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security*, pp. 1575–90. New York: ACM

40. Zhang Y, Steele A, Blanton M. 2013. PICCO: a general-purpose compiler for private distributed computation. In *CCS '13: Proceedings of the 2013 ACM SIGSAC Conference on Computer & Communications Security*, pp. 813–26. New York: ACM

41. Demmler D, Schneider T, Zohner M. 2015. *ABY—a framework for efficient mixed-protocol secure two-party computation.* Paper presented at the Network and Distributed System Security (NDSS) Symposium, San Diego, CA, Feb. 8

42. Liu J, Juuti M, Lu Y, Asokan N. 2017. Oblivious neural network predictions via MiniONN transformations. In *CCS '17: Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, pp. 619–31. New York: ACM

43. Mohassel P, Zhang Y. 2017. SecureML: a system for scalable privacy-preserving machine learning. In *2017 IEEE Symposium on Security and Privacy (S&P)*, pp. 19–38. Piscataway, NJ: IEEE

44. Riazi MS, Weinert C, Tkachenko O, Songhori EM, Schneider T, Koushanfar F. 2018. Chameleon: a hybrid secure computation framework for machine learning applications. In *ASIACCS '18: Proceedings of the 2018 on Asia Conference on Computer and Communications Security*, pp. 707–21. New York: ACM

45. Mohassel P, Rindal P. 2018. ABY3: a mixed protocol framework for machine learning. In *CCS '18: Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security*, pp. 35–52. New York: ACM

46. Makri E, Rotaru D, Vercauteren F, Wagh S. 2021. Rabbit: efficient comparison for secure multi-party computation. In *Financial Cryptography and Data Security: Lecture Notes in Computer Science*, ed. N Borisov, C Diaz, pp. 249–70. Berlin: Springer

47. Cho H, Wu DJ, Berger B. 2018. Secure genome-wide association analysis using multiparty computation. *Nat. Biotechnol.* 36(6):547–51

48. Kamm L, Bogdanov D, Laur S, Vilo J. 2013. A new way to protect privacy in large-scale genome-wide association studies. *Bioinformatics* 29(7):886–93

49. Jagadeesh KA, Wu DJ, Birgmeier JA, Boneh D, Bejerano G. 2017. Deriving genomic diagnoses without revealing patient genomes. *Science* 357(6352):692–95

50. Jha S, Kruger L, Shmatikov V. 2008. Towards practical privacy for genomic computation. In *2008 IEEE Symposium on Security and Privacy (S&P)*, pp. 216–30. Piscataway, NJ: IEEE

51. Bogdanov D, Kamm L, Laur S, Sokk V. 2018. Implementation and evaluation of an algorithm for cryptographically private principal component analysis on genomic data. *Trans. Comput. Biol. Bioinform.* 15(5):1427–32

52. Ma R, Li Y, Li C, Wan F, Hu H, et al. 2020. Secure multiparty computation for privacy-preserving drug discovery. *Bioinformatics* 36(9):2872–80

53. Hie B, Cho H, Berger B. 2018. Realizing private and practical pharmacological collaboration. *Science* 362(6412):347–50

54. von Maltitz M, Ballhausen H, Kaul D, Fleischmann DF, Niyazi M, et al. 2021. A privacy-preserving log-rank test for the Kaplan-Meier estimator with secure multiparty computation: algorithm development and validation. *JMIR Med. Inform.* 9(1):e22158
55. Smajlović H, Shajii A, Berger B, Cho H, Numanagić I. 2023. Sequre: a high-performance framework for secure multiparty computation enables biomedical data sharing. *Genome Biol.* 24(1):5
56. Rivest RL, Shamir A, Adleman L. 1978. A method for obtaining digital signatures and public-key cryptosystems. *Commun. ACM* 21(2):120–26
57. Elgamal T. 1985. A public key cryptosystem and a signature scheme based on discrete logarithms. *IEEE Trans. Inform. Theory* 31(4):469–72
58. Paillier P. 1999. Public-key cryptosystems based on composite degree residuosity classes. In *EUROCRYPT '99: International Conference on the Theory and Applications of Cryptographic Techniques*, ed. J Stern, pp. 223–38. Berlin: Springer
59. Goldwasser S, Micali S. 2019. Probabilistic encryption & how to play mental poker keeping secret all partial information. In *Providing Sound Foundations for Cryptography: On the Work of Shafi Goldwasser and Silvio Micali*, ed. O Goldreich, pp. 203–25. New York: ACM
60. Gentry C. 2009. *A fully homomorphic encryption scheme*. PhD Diss., Stanford Univ., Stanford, CA
61. Gentry C, Halevi S. 2011. Implementing Gentry's fully-homomorphic encryption scheme. In *EUROCRYPT 2011: Advances in Cryptology*, ed. KG Paterson, pp. 129–48. Berlin: Springer
62. Fan J, Vercauteren F. 2012. Somewhat practical fully homomorphic encryption. *Cryptology ePrint Archive*. **https://eprint.iacr.org/2012/144**
63. Brakerski Z, Gentry C, Vaikuntanathan V. 2014. (Leveled) fully homomorphic encryption without bootstrapping. In *ITCS '12: Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, pp. 309–25. New York: ACM
64. Brakerski Z. 2012. Fully homomorphic encryption without modulus switching from classical GapSVP. In *CRYPTO 2012: Advances in Cryptology*, ed. R Safavi-Naini, R Canetti, pp. 868–86. Berlin: Springer
65. Cheon JH, Kim A, Kim M, Song Y. 2017. Homomorphic encryption for arithmetic of approximate numbers. In *Advances in Cryptology—ASIACRYPT 2017: 23rd International Conference on the Theory and Applications of Cryptology and Information Security, Hong Kong, China, December 3–7, 2017, Proceedings, Part I*, pp. 409–37. Cham, Switz.: Springer
66. Regev O. 2009. On lattices, learning with errors, random linear codes, and cryptography. *J. ACM* 56(6):1–40
67. Lyubashevsky V, Peikert C, Regev O. 2010. On ideal lattices and learning with errors over rings. In *Advances in Cryptology—EUROCRYPT 2010: 29th Annual International Conference on the Theory and Applications of Cryptographic Techniques, French Riviera, May 30–June 3, 2010, Proceedings*, ed. H Gilbert, pp. 1–23. Berlin: Springer
68. Bossuat JP, Mouchet C, Troncoso-Pastoriza J, Hubaux JP. 2021. Efficient bootstrapping for approximate homomorphic encryption with non-sparse keys. In *EUROCRYPT 2021: Advances in Cryptology*, ed. A Canteaut, FX Standaert, pp. 587–617. Cham, Switz.: Springer
69. Han K, Ki D. 2020. Better bootstrapping for approximate homomorphic encryption. In *Topics in Cryptology – CT-RSA 2020*, ed. S Jarecki, pp. 364–90. Cham, Switz.: Springer
70. Chillotti I, Gama N, Georgieva M, Izabachène M. 2020. TFHE: fast fully homomorphic encryption over the torus. *J. Cryptol.* 33(1):34–91
71. Viand A, Jattke P, Hithnawi A. 2021. SoK: fully homomorphic encryption compilers. In *2021 IEEE Symposium on Security and Privacy (SP)*, pp. 1092–108. Piscataway, NJ: IEEE
72. Gilad-Bachrach R, Dowlin N, Laine K, Lauter K, Naehrig M, Wernsing J. 2016. Cryptonets: applying neural networks to encrypted data with high throughput and accuracy. *PMLR* 48:201–10
73. Graepel T, Lauter K, Naehrig M. 2012. ML confidential: machine learning on encrypted data. In *ICISC 2012: Information Security and Cryptology*, ed. T Kwon, MK Lee, D Kwon, pp. 1–21. Berlin: Springer
74. Kocabas O, Soyata T. 2020. Towards privacy-preserving medical cloud computing using homomorphic encryption. In *Virtual and Mobile Healthcare: Breakthroughs in Research and Practice*, pp. 93–125. Hershey, PA: IGI Global
75. Bos JW, Lauter K, Naehrig M. 2014. Private predictive analysis on encrypted medical data. *J. Biomed. Inform.* 50:234–43

76. Blatt M, Gusev A, Polyakov Y, Goldwasser S. 2020. Secure large-scale genome-wide association studies using homomorphic encryption. *PNAS* 117(21):11608–13
77. Kim M, Lauter K. 2015. Private genome analysis through homomorphic encryption. *BMC Med. Inform. Decis. Mak.* 15(Suppl. 5):S3
78. Bonte C, Makri E, Ardeshirdavani A, Simm J, Moreau Y, Vercauteren F. 2018. Towards practical privacy-preserving genome-wide association study. *BMC Bioinform.* 19(1):537
79. Lu WJ, Yamada Y, Sakuma J. 2015. Privacy-preserving genome-wide association studies on cloud environment using fully homomorphic encryption. *BMC Med. Inform. Decis. Mak.* 15(Suppl. 5):S1
80. Zhang Y, Dai W, Jiang X, Xiong H, Wang S. 2015. FORESEE: fully outsourced secure genome study based on homomorphic encryption. *BMC Med. Inform. Decis. Mak.* 15(Suppl. 5):S5
81. Leighton AT, Yu YW. 2023. Secure federated Boolean count queries using fully-homomorphic cryptography. bioRxiv 2021.11.10.468090. **https://doi.org/10.1101/2021.11.10.468090**
82. Kantarcioglu M, Jiang W, Liu Y, Malin B. 2008. A cryptographic approach to securely share and query genomic sequences. *IEEE Trans. Inform. Technol. Biomed.* 12(5):606–17
83. Bruekers F, Katzenbeisser S, Kursawe K, Tuyls P. 2008. Privacy-preserving matching of DNA profiles. *Cryptology ePrint Archive*. **https://eprint.iacr.org/2008/203**
84. Ayday E, Raisaro JL, McLaren PJ, Fellay J, Hubaux JP. 2013. *Privacy-preserving computation of disease risk by using genomic, clinical, and environmental data*. Paper presented at the 2013 USENIX Workshop on Health Information Technologies, Washington, DC, Aug. 12
85. Kim M, Harmanci AO, Bossuat JP, Carpov S, Cheon JH, et al. 2021. Ultrafast homomorphic encryption models enable secure outsourcing of genotype imputation. *Cell Syst.* 12(11):1108–20
86. Gürsoy G, Chielle E, Brannon CM, Maniatakos M, Gerstein M. 2022. Privacy-preserving genotype imputation with fully homomorphic encryption. *Cell Syst.* 13(2):173–82
87. Desmedt YG. 1994. Threshold cryptography. *Eur. Trans. Telecommun.* 5(4):449–58
88. Asharov G, Jain A, López-Alt A, Tromer E, Vaikuntanathan V, Wichs D. 2012. Multiparty computation with low communication, computation and interaction via threshold FHE. In *EUROCRYPT 2012: Advances in Cryptology*, ed. D Pointcheval, T Johansson, pp. 483–501. Berlin: Springer
89. Mouchet C, Troncoso-Pastoriza JR, Bossuat JP, Hubaux JP. 2021. Multiparty homomorphic encryption from ring-learning-with-errors. *Proc. Priv. Enhanc. Technol. Symp.* 2021(4):291–311
90. Damgård I, Pastro V, Smart N, Zakarias S. 2012. Multiparty computation from somewhat homomorphic encryption. In *CRYPTO 2012: Advances in Cryptology*, ed. R Safavi-Naini, R Canetti, pp. 643–62. Berlin: Springer
91. Kim T, Kwak H, Lee D, Seo J, Song Y. 2022. Asymptotically faster multi-key homomorphic encryption from homomorphic gadgetc decomposition. *Cryptology ePrint Archive*. **https://eprint.iacr.org/2022/347.pdf**
92. Kwak H, Lee D, Song Y, Wagh S. 2021. A unified framework of homomorphic encryption for multiple parties with non-interactive setup. *Cryptology ePrint Archive*. **https://eprint.iacr.org/2021/1412**
93. Froelicher D, Troncoso-Pastoriza JR, Raisaro JL, Cuendet MA, Sousa JS, et al. 2021. Truly privacy-preserving federated analytics for precision medicine with multiparty homomorphic encryption. *Nat. Commun.* 12(1):5910
94. Cho H, Froelicher D, Chen J, Edupalli M, Pyrgelis A, et al. 2022. Secure and federated genome-wide association studies for biobank-scale datasets. bioRxiv 2022.11.30.518537. **https://doi.org/10.1101/2022.11.30.518537**
95. Froelicher D, Cho H, Edupalli M, Sousa JS, Bossuat J, et al. 2023. Scalable and privacy-preserving federated principal component analysis. In *44th IEEE Symposium on Security and Privacy (SP)*, pp. 1908–25. Piscataway, NJ: IEEE
96. Zheng W, Popa RA, Gonzalez JE, Stoica I. 2019. Helen: maliciously secure coopetitive learning for linear models. In *2019 IEEE Symposium on Security and Privacy (SP)*, pp. 724–38. Piscataway, NJ: IEEE
97. Froelicher D, Troncoso-Pastoriza JR, Pyrgelis A, Sav S, Sousa JS, et al. 2021. Scalable privacy-preserving distributed learning. *Proc. Priv. Enhanc. Technol. Symp.* 2021(2):323–47
98. Sav S, Pyrgelis A, Troncoso-Pastoriza JR, Froelicher D, Bossuat JP, et al. 2021. *POSEIDON: privacy-preserving federated neural network learning*. Paper presented at the 28th Annual Network and Distributed System Security Symposium, online, Feb. 21

99. Yang M, Zhang C, Wang X, Liu X, Li S, et al. 2022. TrustGWAS: a full-process workflow for encrypted GWAS using multi-key homomorphic encryption and pseudorandom number perturbation. *Cell Syst.* 13(9):752–67
100. Sav S, Bossuat JP, Troncoso-Pastoriza JR, Claassen M, Hubaux JP. 2022. Privacy-preserving federated neural network learning for disease-associated cell classification. *Patterns* 3(5):100487
101. Mendelsohn S, Froelicher D, Loginov D, Bernick D, Berger B, Cho H. 2023. sfkit: a web-based toolkit for secure and federated genomic analysis. *Nucleic Acids Res.* 51(W1):W535–41
102. Li W, Kim M, Zhang K, Chen H, Jiang X, Harmanci A. 2023. COLLAGENE enables privacy-aware federated and collaborative genomic data analysis. *Genome Biol.* 24(1):204
103. Sabt M, Achemlal M, Bouabdallah A. 2015. Trusted execution environment: what it is, and what it is not. In *ISPA 2015: The 13th IEEE International Symposium on Parallel and Distributed Processing with Applications*, pp. 57–64. Piscataway, NJ: IEEE
104. Pinto S, Santos N. 2019. Demystifying Arm TrustZone: a comprehensive survey. *ACM Comput. Surv.* 51(6):130
105. Banks AS, Kisiel M, Korsholm P. 2021. Remote attestation: a literature review. arXiv:2105.02466 [cs.CR]
106. Costan V, Devadas S. 2016. Intel SGX explained. *Cryptology ePrint Archive*. **https://eprint.iacr.org/2016/086**
107. Intel Corp. 2022. *Intel® trust domain extensions*. White Pap., Intel Corp., Santa Clara, CA
108. Kaplan D, Powell J, Woller T. 2021. *AMD memory encryption*. White Pap., Adv. Micro Devices, Santa Clara, CA
109. Nertney R. 2023. *Confidential compute on NVIDIA Hopper H100*. White Pap. WP-11459-001, NVIDIA, Santa Clara, CA
110. Borrello P, Kogler A, Schwarzl M, Lipp M, Gruss D, Schwarz M. 2022. *ÆPIC leak: architecturally leaking uninitialized data from the microarchitecture*. Paper presented at the 31st USENIX Security Symposium (USENIX Security 22), Boston, MA, Aug. 10
111. van Schaik S, Seto A, Yurek T, Batori A, AlBassam B, et al. 2022. SoK: SGX.Fail: How stuff gets eXposed. Tech. Rep., Georgia Tech Univ., Atlanta. **https://sgx.fail/files/sgx.fail.pdf**
112. Fei S, Yan Z, Ding W, Xie H. 2021. Security vulnerabilities of SGX and countermeasures: a survey. *ACM Comput. Surv.* 54(6):126
113. Dokmai N, Kockan C, Zhu K, Wang X, Sahinalp SC, Cho H. 2021. Privacy-preserving genotype imputation in a trusted execution environment. *Cell Syst.* 12(10):983–93.e7
114. BeeKeeperAI. 2022. *BeeKeeperAI applies sightless computing technology to pediatric rare disease project*. Press Release, Oct. 19. **https://www.beekeeperai.com/beekeeperai-novartis-pediatric-rare-disease-press-release**
115. Intel Corp. 2021. *Maximum security at the processor level: Intel SGX protects electronic patient record*. Solution Brief, Intel Corp., Santa Clara, CA
116. Pascoal T, Decouchant J, Boutet A, Esteves-Verissimo P. 2021. DyPS: dynamic, private and secure GWAS. *Proc. Priv. Enhanc. Technol.* 2021(2):214–34
117. Kockan C, Zhu K, Dokmai N, Karpov N, Kulekci MO, et al. 2020. Sketching algorithms for genomic data analysis and querying in a secure enclave. *Nat. Methods* 17(3):295–301
118. Widanage C, Liu W, Li J, Chen H, Wang X, et al. 2021. HySec-Flow: privacy-preserving genomic computing with SGX-based big-data analytics framework. *IEEE Int. Conf. Cloud Comput.* 2021:733–43
119. Dwork C, McSherry F, Nissim K, Smith A. 2006. Calibrating noise to sensitivity in private data analysis. In *TCC 2006: Theory of Cryptography* , ed. S Halevi, T Rabin, pp. 265–84. Berlin: Springer
120. Dwork C, Kenthapadi K, McSherry F, Mironov I, Naor M. 2006. Our data, ourselves: privacy via distributed noise generation. In *EUROCRYPT 2006: Advances in Cryptology*, ed. S Vaudenay, pp. 486–503. Berlin: Springer
121. Dwork C, Rothblum GN. 2016. Concentrated differential privacy. arXiv:1603.01887 [cs.DS]
122. Bun M, Steinke T. 2016. Concentrated differential privacy: simplifications, extensions, and lower bounds. In *TCC 2016: Theory of Cryptography*, ed. M Hirt, A Smith, pp. 635–58. Berlin: Springer
123. Mironov I. 2017. Rényi differential privacy. In *IEEE 30th Computer Security Foundations Symposium*, pp. 263–75. Piscataway, NJ: IEEE

124. Chaudhuri K, Monteleoni C, Sarwate AD. 2011. Differentially private empirical risk minimization. *J. Mach. Learn. Res.* 12(3):1069–109

125. Iyengar R, Near JP, Song D, Thakkar O, Thakurta A, Wang L. 2019. Towards practical differentially private convex optimization. In *2019 IEEE Symposium on Security and Privacy (SP)*, pp. 299–316. Piscataway, NJ: IEEE

126. Bassily R, Smith A, Thakurta A. 2014. Private empirical risk minimization: efficient algorithms and tight error bounds. In *2014 IEEE 55th Annual Symposium on Foundations of Computer Science*, pp. 464–73. Piscataway, NJ: IEEE

127. Abadi M, Chu A, Goodfellow I, McMahan HB, Mironov I, et al. 2016. Deep learning with differential privacy. In *CCS '16: Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, pp. 308–18. New York: ACM

128. Nissim K, Raskhodnikova S, Smith A. 2007. Smooth sensitivity and sampling in private data analysis. In *STOC '07: Proceedings of the Thirty-Ninth Annual ACM Symposium on Theory of Computing*, pp. 75–84. New York: ACM

129. Warner SL. 1965. Randomized response: a survey technique for eliminating evasive answer bias. *J. Am. Stat. Assoc.* 60(309):63–69

130. Dwork C, Naor M, Reingold O, Rothblum GN, Vadhan S. 2009. On the complexity of differentially private data release: efficient algorithms and hardness results. In *STOC '09: Proceedings of the Forty-First Annual ACM Symposium on Theory of Computing*, pp. 381–90. New York: ACM

131. Kairouz P, Bonawitz K, Ramage D. 2016. Discrete distribution estimation under local privacy. *PMLR* 48:2436–44

132. Erlingsson Ú, Pihur V, Korolova A. 2014. RAPPOR: randomized aggregatable privacy-preserving ordinal response. In *CCS '14: Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security*, pp. 1054–67. New York: ACM

133. Differential Privacy Team. 2017. *Learning with privacy at scale*. Mach. Learn. Res., Apple, Cupertino, CA. **https://machinelearning.apple.com/research/learning-with-privacy-at-scale**

134. Ding B, Kulkarni J, Yekhanin S. 2017. Collecting telemetry data privately. In *NIPS '17: Proceedings of the 31st International Conference on Neural Information Processing Systems*, ed. U von Luxburg, pp. 3574–83. Red Hook, NY: Curran Assoc.

135. Abowd J, Kifer D, Garfinkel SL, Machanavajjhala A. 2019. *Census TopDown: differentially private data, incremental schemas, and consistency with public knowledge*. Tech. Pap., US Census Bureau, Silver Hill, MD

136. Uhlerop C, Slavković A, Fienberg SE. 2013. Privacy-preserving data sharing for genome-wide association studies. *J. Priv. Confid.* 5(1):137–66

137. Yu F, Fienberg SE, Slavković AB, Uhler C. 2014. Scalable privacy-preserving data sharing methodology for genome-wide association studies. *J. Biomed. Inform.* 50:133–41

138. Yu F, Rybar M, Uhler C, Fienberg SE. 2014. Differentially-private logistic regression for detecting multiple-SNP association in GWAS databases. In *PSD 2014: Privacy in Statistical Databases*, ed. J Domingo-Ferrer, pp. 170–84. Berlin: Springer

139. Johnson A, Shmatikov V. 2013. Privacy-preserving data exploration in genome-wide association studies. In *KDD '13: Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1079–87. New York: ACM

140. Simmons S, Berger B. 2016. Realizing privacy preserving genome-wide association studies. *Bioinformatics* 32(9):1293–300

141. Simmons S, Sahinalp C, Berger B. 2016. Enabling privacy-preserving GWASs in heterogeneous human populations. *Cell Syst.* 3(1):54–61

142. Wang S, Mohammed N, Chen R. 2014. Differentially private genome data dissemination through top-down specialization. *BMC Med. Inform. Decis. Mak.* 14(Suppl. 1):S2

143. Beaulieu-Jones BK, Wu ZS, Williams C, Lee R, Bhavnani SP, et al. 2019. Privacy-preserving generative deep neural networks support clinical data sharing. *Circ. Cardiovasc. Qual. Outcomes* 12(7):e005122

144. Mohammed N, Jiang X, Chen R, Fung BC, Ohno-Machado L. 2013. Privacy-preserving heterogeneous health data sharing. *J. Am. Med. Inform. Assoc.* 20(3):462–69

145. Cho H, Simmons S, Kim R, Berger B. 2020. Privacy-preserving biomedical database queries with optimal privacy-utility trade-offs. *Cell Syst.* 10(5):408–16

146. Vinterbo SA, Sarwate AD, Boxwala AA. 2012. Protecting count queries in study design. *J. Am. Med. Inform. Assoc.* 19(5):750–57

147. Wei J, Lin Y, Yao X, Zhang J, Liu X. 2020. Differential privacy-based genetic matching in personalized medicine. *IEEE Trans. Emerg. Top. Comput.* 9(3):1109–25

148. Field E, Dyda A, Lau C. 2021. COVID-19 real-time information system for preparedness and epidemic response (CRISPER). *Med. J. Aust.* 214(8):386–86.e1

149. Liu X, Zhou P, Qiu T, Wu DO. 2020. Blockchain-enabled contextual online learning under local differential privacy for coronary heart disease diagnosis in mobile edge computing. *IEEE J. Biomed. Health Informat.* 24(8):2177–88

150. Kairouz P, McMahan HB, Avent B, Bellet A, Bennis M, et al. 2021. Advances and open problems in federated learning. *Found. Trends Mach. Learn.* 14(1–2):1–210

151. McMahan B, Moore E, Ramage D, Hampson S, Agüera y Arcas B. 2017. Communication-efficient learning of deep networks from decentralized data. *PMLR* 54:1273–82

152. Li T, Sanjabi M, Beirami A, Smith V. 2020. *Fair resource allocation in federated learning*. Poster presented at ICLR 2020: International Conference on Learning Representations, Addis Ababa, Ethiopia, Apr. 30

153. Reddi SJ, Charles Z, Zaheer M, Garrett Z, Rush K, et al. 2020. *Adaptive federated optimization*. Paper presented at ICLR 2021: International Conference on Learning Representations, Vienna, May 4

154. Wang H, Yurochkin M, Sun Y, Papailiopoulos D, Khazaeni Y. 2020. *Federated learning with matched averaging*. Paper presented at ICLR 2020: International Conference on Learning Representations, online, Apr. 26

155. Hegedűs I, Danner G, Jelasity M. 2019. Gossip learning as a decentralized alternative to federated learning. In *Distributed Applications and Interoperable Systems: 19th IFIP WG 6.1 International Conference, DAIS 2019, Held as Part of the 14th International Federated Conference on Distributed Computing Techniques, DisCoTec 2019, Kongens Lyngby, Denmark, June 17–21, 2019, Proceedings*, ed. J Pereira, L Ricci, pp. 74–90. Cham, Switz.: Springer

156. Tan AZ, Yu H, Cui L, Yang Q. 2023. Towards personalized federated learning. *IEEE Trans. Neural Netw. Learn. Syst.* 34(12):9587–603

157. Achituve I, Shamsian A, Navon A, Chechik G, Fetaya E. 2021. Personalized federated learning with Gaussian processes. *Adv. Neural Inform. Proc. Syst.* 34:8392–406

158. Wang S, Tuor T, Salonidis T, Leung KK, Makaya C, et al. 2019. Adaptive federated learning in resource constrained edge computing systems. *IEEE J. Sel. Areas Commun.* 37(6):1205–21

159. Zhao Y, Li M, Lai L, Suda N, Civin D, Chandra V. 2018. Federated learning with non-IID data. arXiv:1806.00582 [cs.LG]

160. Li T, Hu S, Beirami A, Smith V. 2021. Ditto: fair and robust federated learning through personalization. *PMLR* 139:6357–68

161. Michieli U, Ozay M. 2021. Are all users treated fairly in federated learning systems? In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2318–22. Nashville, TN: IEEE

162. Zhang DY, Kou Z, Wang D. 2020. FairFL: a fair federated learning approach to reducing demographic bias in privacy-sensitive classification models. In *Proceedings of the 2020 IEEE International Conference on Big Data*, ed. X Wu, C Jermaine, L Xiong, O Kotevska, S Lu, et al., pp. 1051–60. Piscataway, NJ: IEEE

163. So J, Ali RE, Güler B, Jiao J, Avestimehr AS. 2023. Securing secure aggregation: mitigating multi-round privacy leakage in federated learning. In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence*, pp. 9864–73. Washington, DC: AAAI

164. Geiping J, Bauermeister H, Dröge H, Moeller M. 2020. Inverting gradients—how easy is it to break privacy in federated learning? *Adv. Neural Inform. Proc. Syst.* 33:16937–47

165. Huang Y, Gupta S, Song Z, Li K, Arora S. 2021. Evaluating gradient inversion attacks and defenses in federated learning. *Adv. Neural Inform. Proc. Syst.* 34:7232–41

166. Al Mallah R, Lopez D, Badu-Marfo G, Farooq B. 2021. Untargeted poisoning attack detection in federated learning via behavior attestation. *IEEE Access* 11:125064–79

167. Tolpegin V, Truex S, Gursoy ME, Liu L. 2020. Data poisoning attacks against federated learning systems. In *Computer Security—ESORICS 2020: 25th European Symposium on Research in Computer Security, ESORICS 2020, Guildford, UK, September 14–18, 2020, Proceedings, Part I*, ed. L Chen, N Li, K Liang, S Schneider, pp. 480–501. Cham, Switz.: Springer

168. Hu R, Guo Y, Li H, Pei Q, Gong Y. 2020. Personalized federated learning with differential privacy. *IEEE Internet Things J.* 7(10):9530–39

169. Noble M, Bellet A, Dieuleveut A. 2022. Differentially private federated learning on heterogeneous data. *PMLR* 151:10110–45

170. Truex S, Liu L, Chow KH, Gursoy ME, Wei W. 2020. LDP-Fed: federated learning with local differential privacy. In *EdgeSys '20: Proceedings of the Third ACM International Workshop on Edge Systems, Analytics and Networking*, pp. 61–66. New York: ACM

171. Grammenos A, Mendoza Smith R, Crowcroft J, Mascolo C. 2020. Federated principal component analysis. *Adv. Neural Inform. Proc. Syst.* 33:6453–64

172. Mansour Y, Mohri M, Ro J, Suresh AT. 2020. Three approaches for personalization with applications to federated learning. arXiv:2002.10619 [cs.LG]

173. Chen Y, Qin X, Wang J, Yu C, Gao W. 2020. FedHealth: a federated transfer learning framework for wearable healthcare. *IEEE Intel. Syst.* 35(4):83–93

174. Pati S, Baid U, Edwards B, Sheller M, Wang SH, et al. 2022. Federated learning enables big data for rare cancer boundary detection. *Nat. Commun.* 13(1):7346

175. Darzidehkalani E, Ghasemi-Rad M, van Ooijen P. 2022. Federated learning in medical imaging: part I: toward multicentral health care ecosystems. *J. Am. Coll. Radiol.* 19(8):969–74

176. Ng D, Lan X, Yao MM-S, Chan WP, Feng M. 2021. Federated learning: a collaborative effort to achieve better medical imaging models for individual sites that have small labelled datasets. *Quant. Imaging Med. Surg.* 11(2):852–57

177. Sarma KV, Harmon S, Sanford T, Roth HR, Xu Z, et al. 2021. Federated learning improves site performance in multicenter deep learning without data sharing. *J. Am. Med. Inform. Assoc.* 28(6):1259–64

178. Kaissis G, Ziller A, Passerat-Palmbach J, Ryffel T, Usynin D, et al. 2021. End-to-end privacy preserving deep learning on multi-institutional medical imaging. *Nat. Mach. Intel.* 3(6):473–84

179. Vaid A, Jaladanki SK, Xu J, Teng S, Kumar A, et al. 2021. Federated learning of electronic health records to improve mortality prediction in hospitalized patients with COVID-19: machine learning approach. *JMIR Med. Inform.* 9(1):e24207

180. Brisimi TS, Chen R, Mela T, Olshevsky A, Paschalidis IC, Shi W. 2018. Federated learning of predictive models from federated electronic health records. *Int. J. Med. Inform.* 112:59–67

181. Liu D, Dligach D, Miller T. 2019. Two-stage federated phenotyping and patient representation learning. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, ed. D Demner-Fushman, K Bretonnel Cohen, S Ananiadou, J Tsujii, pp. 283–91. Florence, Italy: Assoc. Comp. Linguist.

182. Paulik M, Seigel M, Mason H, Telaar D, Kluivers J, et al. 2021. Federated evaluation and tuning for on-device personalization: system design & applications. arXiv:2102.08503 [cs.LG]

183. Wu Q, Chen X, Zhou Z, Zhang J. 2020. FedHome: cloud-edge based personalized federated learning for in-home health monitoring. *IEEE Trans. Mobile Comput.* 21(8):2818–32

184. Ghosh S, Ghosh SK. 2023. FEEL: federated learning framework for elderly healthcare using Edge-IoMT. *IEEE Trans. Comput. Soc. Syst.* 10:1800–9

185. Chor B, Kushilevitz E, Goldreich O, Sudan M. 1998. Private information retrieval. *J. ACM* 45(6):965–81

186. Kushilevitz E, Ostrovsky R. 1997. Replication is not needed: single database, computationally-private information retrieval. In *Proceedings of the 38th Annual Symposium on Foundations of Computer Science*, pp. 364–73. Piscataway, NJ: IEEE

187. Beimel A, Ishai Y, Malkin T. 2000. Reducing the servers computation in private information retrieval: PIR with preprocessing. In *CRYPTO 2000: Advances in Cryptology*, ed. M Bellare, pp. 55–73. Berlin: Springer

188. Corrigan-Gibbs H, Kogan D. 2020. Private information retrieval with sublinear online time. In *EUROCRYPT 2020: Advances in Cryptology*, ed. A Canteaut, Y Ishai, pp. 44–75. Cham, Switz.: Springer

189. Melchor CA, Barrier J, Fousse L, Killijian MO. 2016. XPIR: private information retrieval for everyone. *Proc. Priv. Enhanc. Technol.* 2016:155–74

190. Davidson A, Pestana G, Celi S. 2022. FrodoPIR: simple, scalable, single-server private information retrieval. *Cryptology ePrint Archive*. **https://eprint.iacr.org/2022/981**

191. Menon SJ, Wu DJ. 2022. SPIRAL: fast, high-rate single-server PIR via FHE composition. In *2022 IEEE Symposium on Security and Privacy (SP)*, pp. 930–47. Piscataway, NJ: IEEE

192. Henzinger A, Hong MM, Corrigan-Gibbs H, Meiklejohn S, Vaikuntanathan V. 2023. One server for the price of two: simple and fast single-server private information retrieval. In *32nd USENIX Security Symposium*. Berkeley, CA: USENIX. **https://www.usenix.org/system/files/sec23summer_27-henzinger-prepub.pdf**
193. Chor B, Gilboa N, Naor M. 1997. *Private information retrieval by keywords*. Tech. Rep. TR CS0917, Dep. Comput. Sci., Technion, Haifa, Israel
194. Patel S, Seo JY, Yeo K. 2023. Don't be dense: efficient keyword PIR for sparse databases. *32nd USENIX Security Symposium*. Berkeley, CA: USENIX. **https://www.usenix.org/system/files/sec23fall-prepub-392-patel.pdf**
195. Ishai Y, Kushilevitz E, Ostrovsky R, Sahai A. 2004. Batch codes and their applications. In *STOC '04: Proceedings of the Thirty-Sixth Annual ACM Symposium on Theory of Computing*, pp. 262–71. New York: ACM
196. Angel S, Chen H, Laine K, Setty S. 2018. PIR with compressed queries and amortized query processing. In *2018 IEEE Symposium on Security and Privacy (SP)*, pp. 962–79. Piscataway, NJ: IEEE
197. Sousa JS, Lefebvre C, Huang Z, Raisaro JL, Aguilar-Melchor C, et al. 2017. Efficient and secure outsourcing of genomic data storage. *BMC Med. Genom.* 10(Suppl. 2):46
198. Çetin GS, Chen H, Laine K, Lauter K, Rindal P, Xia Y. 2017. Private queries on encrypted genomic data. *BMC Med. Genom.* 10(2):45
199. Freedman MJ, Nissim K, Pinkas B. 2004. Efficient private matching and set intersection. In *EUROCRYPT 2004: Advances in Cryptology*, ed. C Cachin, JL Camenisch, pp. 1–19. Berlin: Springer
200. Pinkas B, Rosulek M, Trieu N, Yanai A. 2019. SpOT-light: lightweight private set intersection from sparse OT extension. In *CRYPTO 2019: Advances in Cryptology*, ed. A Boldyreva, D Micciancio, pp. 401–31. Cham, Switz.: Springer
201. Chase M, Miao P. 2020. Private set intersection in the internet setting from lightweight oblivious PRF. In *CRYPTO 2020: Advances in Cryptology*, ed. D Micciancio, T Ristenpart, pp. 34–63. Cham, Switz.: Springer
202. Baldi P, Baronio R, De Cristofaro E, Gasti P, Tsudik G. 2011. Countering GATTACA: efficient and secure testing of fully-sequenced human genomes. In *CCS '11: Proceedings of the 18th ACM Conference on Computer and Communications Security*, pp. 691–702. New York: ACM
203. Agrawal R, Evfimievski A, Srikant R. 2003. Information sharing across private databases. In *SIGMOD '03: Proceedings of the 2003 ACM SIGMOD International Conference on Management of Data*, pp. 86–97. New York: ACM
204. De Cristofaro E, Gasti P, Tsudik G. 2012. Fast and private computation of cardinality of set intersection and union. In *CANS 2012: Cryptology and Network Security*, ed. J Pieprzyk, AR Sadeghi, M Manulis, pp. 218–31. Berlin: Springer
205. Wang XS, Huang Y, Zhao Y, Tang H, Wang X, Bu D. 2015. Efficient genome-wide, privacy-preserving similar patient query based on private edit distance. In *CCS '15: Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*, pp. 492–503. New York: ACM
206. Goldwasser S, Micali S, Rackoff C. 1985. The knowledge complexity of interactive proof-systems. In *STOC '85: Proceedings of the Seventeenth Annual ACM Symposium on Theory of Computing*, pp. 291–304. New York: ACM
207. Diffie W, Hellman M. 1976. New directions in cryptography. *IEEE Trans. Inform. Theory* 22(6):644–54
208. Goldreich O, Micali S, Wigderson A. 1987. How to play ANY mental game. In *STOC '87: Proceedings of the Nineteenth Annual ACM Symposium on Theory of Computing*, pp. 218–29. New York: ACM
209. Goldreich O, Micali S, Wigderson A. 1991. Proofs that yield nothing but their validity or all languages in NP have zero-knowledge proof systems. *J. ACM* 38(3):690–728
210. Parno B, Howell J, Gentry C, Raykova M. 2016. Pinocchio: nearly practical verifiable computation. *Commun. ACM* 59(2):103–12
211. Ben-Sasson E, Bentov I, Horesh Y, Riabzev M. 2018. Scalable, transparent, and post-quantum secure computational integrity. *Cryptology ePrint Archive*. **https://eprint.iacr.org/2018/046.pdf**
212. Bünz B, Bootle J, Boneh D, Poelstra A, Wuille P, Maxwell G. 2018. Bulletproofs: Short proofs for confidential transactions and more. In *2018 IEEE Symposium on Security and Privacy (SP)*, pp. 315–34. Piscataway, NJ: IEEE

213. Xie T, Zhang Y, Song D. 2022. Orion: zero knowledge proof with linear prover time. In *CRYPTO 2022: Advances in Cryptology*, ed. Y Dodis, T Shrimpton, pp. 299–328. Cham, Switz.: Springer

214. Froelicher D, Egger P, Sousa JS, Raisaro JL, Huang Z, et al. 2017. UnLynx: a decentralized system for privacy-conscious data sharing. *Proc. Privacy Enhanc. Technol.* 2017(4):232–50

215. Camenisch J, Stadler M. 1997. *Proof systems for general statements about discrete logarithms*. Tech. Rep. 260, Dept. Comp. Sci., ETH Zurich, Zurich. **https://crypto.ethz.ch/publications/files/CamSta97b.pdf**

216. Chatel S, Pyrgelis A, Troncoso-Pastoriza JR, Hubaux JP. 2021. Privacy and integrity preserving computations with CRISP. In *30th USENIX Security Symposium*, pp. 2111–28. Berkeley, CA: USENIX

217. Chase M, Derler D, Goldfeder S, Orlandi C, Ramacher S, et al. 2017. Post-quantum zero-knowledge and signatures from symmetric-key primitives. In *CCS '17: Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, pp. 1825–42. New York: ACM

218. Ishai Y, Kushilevitz E, Ostrovsky R, Sahai A. 2009. Zero-knowledge proofs from secure multiparty computation. *SIAM J. Comput.* 39(3):1121–52

219. Kuo TT, Kim HE, Ohno-Machado L. 2017. Blockchain distributed ledger technologies for biomedical and health care applications. *J. Am. Med. Inform. Assoc.* 24(6):1211–20

220. Esmaeilzadeh P, Mirzaei T. 2019. The potential of blockchain technology for health information exchange: experimental study from patients' perspectives. *J. Med. Internet Res.* 21(6):e14184

221. Grishin D, Raisaro JL, Troncoso-Pastoriza JR, Obbad K, Quinn K, et al. 2021. Citizen-centered, auditable and privacy-preserving population genomics. *Nat. Comput. Sci.* 1(3):192–98

222. Yan C, Yan Y, Wan Z, Zhang Z, Omberg L, et al. 2022. A multifaceted benchmarking of synthetic electronic health record generation models. *Nat. Commun.* 13(1):7609

223. Kazerouni A, Aghdam EK, Heidari M, Azad R, Fayyaz M, et al. 2023. Diffusion models in medical imaging: a comprehensive survey. *Med. Image Anal.* 88:102846

224. Jeon M, Park H, Kim HJ, Morley M, Cho H. 2022. *k*-SALSA: *k*-anonymous synthetic averaging of retinal images via local style alignment. In *ECCV 2022: Computer Vision*, ed. S Avidan, G Brostow, M Cisse, GM Farinella, T Hassner, pp. 661–78. Cham, Switz.: Springer

225. Zhang Z, Yan C, Lasko TA, Sun J, Malin BA. 2021. SynTEG: a framework for temporal structured electronic health data simulation. *J. Am. Med. Inform. Assoc.* 28(3):596–604

226. Zhang Z, Yan C, Malin BA. 2022. Membership inference attacks against synthetic health data. *J. Biomed. Inform.* 125:103977

227. Carlini N, Hayes J, Nasr M, Jagielski M, Sehwag V, et al. 2023. Extracting training data from diffusion models. In *32nd USENIX Security Symposium*, pp. 5253–70. Berkeley, CA: USENIX

228. Torkzadehmahani R, Kairouz P, Paten B. 2019. DP-CGAN: differentially private synthetic data and label generation. In *Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 98–104. Piscataway, NJ: IEEE

229. Sadhuka S, Fridman D, Berger B, Cho H. 2023. Assessing transcriptomic reidentification risks using discriminative sequence models. *Genome Res.* 33(7):1101–12

230. Gürsoy G, Li T, Liu S, Ni E, Brannon CM, Gerstein MB. 2022. Functional genomics data: privacy risk assessment and technological mitigation. *Nat. Rev. Genet.* 23(4):245–58

231. Schadt EE, Woo S, Hao K. 2012. Bayesian method to predict individual SNP genotypes from gene expression data. *Nat. Genet.* 44(5):603–8

232. Hill AC, Guo C, Litkowski EM, Manichaikul AW, Yu B, et al. 2023. Large scale proteomic studies create novel privacy considerations. *Sci. Rep.* 13(1):9254

233. Li H, Wu J, Gao Y, Shi Y. 2016. Examining individuals' adoption of healthcare wearable devices: an empirical study from privacy calculus perspective. *Int. J. Med. Inform.* 88:8–17

234. Nasr M, Carlini N, Hayase J, Jagielski M, Cooper AF, et al. 2023. Scalable extraction of training data from (production) language models. arXiv:2311.17035 [cs.LG]

235. Guo J, Clayton EW, Kantarcioglu M, Vorobeychik Y, Wooders M, et al. 2023. A game theoretic approach to balance privacy risks and familial benefits. *Sci. Rep.* 13(1):6932

236. Xia W, Liu Y, Wan Z, Vorobeychik Y, Kantacioglu M, et al. 2021. Enabling realistic health data re-identification risk assessment through adversarial modeling. *J. Am. Med. Inform. Assoc.* 28(4):744–52

237. Berrang P, Humbert M, Zhang Y, Lehmann I, Eils R, Backes M. 2018. Dissecting privacy risks in biomedical data. In *2018 IEEE European Symposium on Security and Privacy (EuroS&P)*, pp. 62–76. Piscataway, NJ: IEEE

238. Mironov I. 2012. On significance of the least significant bits for differential privacy. In *CCS '12: Proceedings of the 2012 ACM Conference on Computer and Communications Security*, pp. 650–61. New York: ACM

239. Liu C, Chakraborty S, Mittal P. 2016. Dependence makes you vulnerable: differential privacy under dependent tuples. In *Network and Distributed System Security Symposium 2016*, pp. 21–24. Red Hook, NY: Curran Assoc.

240. Chen IY, Pierson E, Rose S, Joshi S, Ferryman K, Ghassemi M. 2021. Ethical machine learning in healthcare. *Annu. Rev. Biomed. Data Sci.* 4:123–44

241. Ding Y, Hou K, Xu Z, Pimplaskar A, Petter E, et al. 2023. Polygenic scoring accuracy varies across the genetic ancestry continuum. *Nature* 618:774–81

242. Movva R, Shanmugam D, Hou K, Pathak P, Guttag J, et al. 2023. Coarse race data conceals disparities in clinical risk score performance. arXiv:2304.09270 [cs.CY]

243. Bak M, Madai VI, Fritzsche MC, Mayrhofer MT, McLennan S. 2022. You can't have AI both ways: balancing health data privacy and access fairly. *Front. Genet.* 13:1490

244. Seastedt KP, Schwab P, O'Brien Z, Wakida E, Herrera K, et al. 2022. Global healthcare fairness: We should be sharing more, not less, data. *PLOS Digit. Health* 1(10):e0000102

245. Xiao Y, Lim S, Pollard TJ, Ghassemi M. 2023. In the name of fairness: assessing the bias in clinical record de-identification. In *FACCT '23: Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pp. 123–37. New York: ACM

246. Suriyakumar VM, Papernot N, Goldenberg A, Ghassemi M. 2021. Chasing your long tails: differentially private prediction in health care settings. In *FACCT '21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp. 723–34. New York: ACM

247. Mayer RC, Davis JH, Schoorman FD. 1995. An integrative model of organizational trust. *Acad. Manag. Rev.* 20(3):709–34