International Journal of Innovative Research in Computer Science & Technology (IJIRCST)
ISSN: 2347-5552, Volume-10, Issue-6, November 2022
https://doi.org/10.55524/ijircst.2022.10.6.22
Article ID IRPV1076, Pages 142-148
www.ijircst.org

Big Data Privacy in Biomedical Research

Mohammad Suhail[1], and Jasdeep Singh[2]

[1]M. Tech Scholar, Department of Computer Science & Engineering, RIMT University, Mandi Gobindgarh, Punjab, India
[2]Assistant Professor, Department of Computer Science & Engineering, RIMT University, Mandi Gobindgarh, Punjab, India

ABSTRACT- The examination of patient data, which may contain personally identifiable information, is a common part of biomedical research. If these data are misused, it could result in the disclosure of private patient information, which would put the patients' right to privacy at risk. The challenge of protecting the privacy of patients in an era dominated by big data has garnered a growing amount of attention in recent years. There have been a lot of different privacy approaches created to protect against different attack models. In the context of research in biomedicine, this publication provides a review of pertinent subjects. It is discussed how technology can protect privacy, particularly in relation to record linking, synthetic data production, and the privacy of genomic data. In addition to this, we conduct an analysis of the ethical implications of the privacy of big data in biomedicine and we emphasise the obstacles that lie ahead for future research pathways aimed at strengthening data privacy in biomedical investigations. Both of these topics are covered in detail throughout this article. After the paper was first published, it was highlighted in the publication Biomedical Research.

KEYWORDS- Data Privacy, Biomedical Research, Data Security, Bioethics, Genome Analysis

I. INTRODUCTION

The Health Information Technology for Economic and Clinical Health Act (HITECH Act) [1] has mandated the adoption of electronic health records (EHRs) in the United States in order to improve the quality of health care, and as of January 2015, 83% of office-based physicians had adopted EHRs. The widespread adoption of electronic health record systems has made it possible for medical professionals and researchers to generate and compile large-scale phenotypic data from patients suffering from a variety of diseases. This is made possible thanks to the fact that it is now possible to do so. In addition to the information that may be accessed via EHR systems, recent advancements in sequencing technology have made human genomic data substantially more accessible and affordable. A national cohort that will encompass one million Americans and have their genetic data sequenced is going to be established as part of the Precision Medicine Initiative, which was just just announced by President Obama. This initiative was just recently announced by President Obama. The United States of America will be the setting for this event. In order to accomplish this objective, it will first merge genetic data and EHR data from networks that have already been formed, and then it will recruit additional participants to take part in the study[4]. These most recent developments make the field of big data science possible and have the potential to significantly accelerate the process of locating new findings in the biomedical field. In addition, the field of big data science has the potential to significantly accelerate the process of discovering new treatments. On the other hand, the ever-increasing amount of biological data, which includes a significant amount of personal information about patients, makes it more difficult than it has ever been to protect the patients' right to personal privacy. This is because the data contains a significant amount of private information about patients. These numbers need to be examined very carefully. ly protected, as failure to do so could result in the disclosure of information and a breach of patients' privacy, which would have a detrimental impact on patients and may have serious repercussions (e.g., discrimi- nation for employment, insurance, or education [2]).

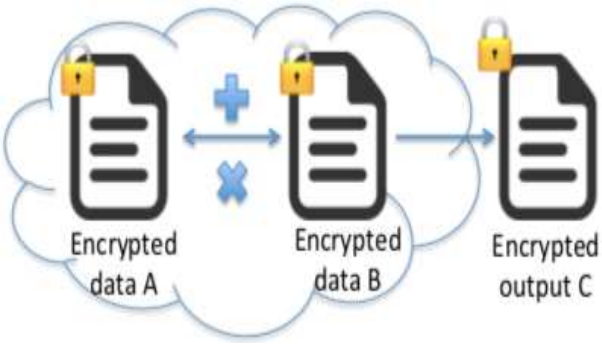

Figure 1: Graphical illustration of the homomorphic encryption algorithm, which allows for computation to be performed on encrypted data while still producing encrypted output

This type of encryption makes it possible to perform homomorphic encryption.
The second approach, known as Safe Harbor, calls for the elimination of a list containing 18 identifiers [3]. It would appear that the Safe Harbor strategy is more beneficial than the Expert Determination strategy when put into practice. The reason for this is that Safe Harbor is more user-friendly. There are several disagreements regarding these HIPAA privacy standards [5] [8], despite the fact that the method is still the most common one in practise. [5] [6] [7] [8] Some people have the opinion that the safeguards against the de-identification of data are not strong enough [5]. The privacy legislation that are currently in effect do not cover either longitudinal data or transactional data, both of which can be used to re-identify a person;

nevertheless, none of these types of data are addressed. In spite of the fact that HIPPA makes a point of mentioning the protection of biometric information such as fingerprints and voice prints, this coverage does not extend to personal genetic data. Others contend that privacy precautions will make it more difficult to do scientific research, and that making them a reality will make it more difficult to conduct significant biomedical investigations that rely on suppressed qualities. One illustration of this would be fine-grained geriatric research carried out in places with less than 20,000 people living in them [9]. These studies would involve participants who were at least 89 years old. One individual has expressed concern that the privacy restriction will make it more difficult to successfully use computerised health records in research [5]. Any data access policy will, in practise, require some form of implicit compromise between the potential threats to an individual's privacy and the benefits that can be gained from the data [10]. In point of fact, the vast majority of owners of clinical data opt for a solution that is somewhere in the middle. To accomplish this, they pick one of the two approaches listed below: (1) modifying data in such a manner that it becomes more difficult to link information to a specific individual, or (2) limiting the amount of information that is made available to the general public. Both of these methods are considered to be forms of information concealment. A viable solution needs to take into account both the context of the application and the expected background knowledge of the attackers in order to find the best possible compromise. Only then can the ideal compromise be found. Only after that would it be possible to achieve the highest possible level of safety[11]. In order to determine the scope of this endeavor, we decided to investigate associated privacy preserving strategies for a select few pertinent and applicable topics in the field of biomedical research. We have high hopes that our discoveries would prove to be beneficial. These locations were chosen due of the possible influence they could have [12]. The following is a list of topics that will be discussed as part of this event: The first is record linkage, the second is distributed data analysis, the third is the development of synthetic data, and the fourth is secure genome analyses. When we change our focus to privacy protection technologies, electronic health records (EHR) and genomic data will both be at the forefront of our attention as we make this transition. The following outline describes the organisation of the remaining parts of this work: In order to get started, Section 2 is going to give the historical context of the problem with data privacy in terms of the ethical and privacy considerations that are involved. Second, in Section 3, we will discuss the processes that are carried out in order to safeguard the confidentiality of the data. The challenges that were encountered, as well as some possible solutions, are spoken about in the fourth section. The verdict will be presented in the fifth and final portion of this article. As seen in Figure 1, HME permits direct calculation over encrypted data using multiplication and addition, with the output encrypted under the same encryption key.

II. PRIVACY PRESERVING METHODS

Here, we'll take a look at privacy-protecting strategies for four distinct types of data: genomic data, synthetic data, EHR patient linkage, and the generation of synthetic data all play important roles in this study.

A. EHR Patient Linkage

These days, biomedical information systems are capable of collecting, storing, and processing vast quantities of data. In spite of the efforts that have been made to manage these information systems, the data that pertain to healthcare are typically disorganized, redundant, prone to error, and heterogeneous. This makes it difficult to obtain information that is meaningful. Record linkage, also known as duplicate detection [35] or entity resolution [36], is a crucial stage in the process of carrying out biomedical research. This step is also known as "entity resolution." It does this by searching multiple different data sources for records that refer to the same real-world thing and then identifying those records when it finds them.

B. The Privacy-Preserving Record Linkage Process

In addition to their privacy approach, PPRL solutions make use of linking procedures that involve two parties and three parties respectively. In the first approach, data owners use encrypted messaging to determine matched pairs on their own directly. [41], [42] make advantage of SMC to compute the set of matching records in a stealthy manner. In the three-party arrangement, a third party is responsible for matching the records that are held by the original data owners. The majority of modern PPRL solutions have a bias toward mistrusting third parties, even when those third parties are reputable institutions such as the National Center for Health Statistics [43]. One protocol party may threaten privacy in both cases. PPRL's "semi-honest" threat model is popular.

C. Secure Transformation Using Bloom Filters

Secure transformations decompose record string attributes into n-grams. The original records are mapped into Bloom filters via these functions as well as the hash function [44]. Bloom filters, often known as BFs, are able to accurately represent the original string thanks to their utilisation of bit arrays. Hash functions produce a probabilistic map that corresponds each letter in the text to a bit in the array. This map is called the hash value. Finding patterns can be accomplished with the help of this map. In this operation, all of the bits in the array are first set to zero; after that, the only bits that are set to one are the ones that were identified by hashing the original string's grammes. This method produces a field-level binary forest (FBF) by determining the degree to which individual records are comparable to one another using set-to-set distance metrics like the Dice coefficient. The PPRL methods that are based on this encoding produce good usefulness; however, the FBF representation of the original string property could potentially reveal some information to an adversary [45]. (for instance, attacks that vary in accordance with their frequency) In order to make BF-based PPRL strategies safer, Durham et al. [46] suggested fusing together numerous FBFs representing each record attribute into a single composite BF for each record. This would result in a single composite BF for the entire record. Because of the composite structure, it will be more difficult for the adversary to determine the original records based on the single bit values by using the frequency information that is included within the grammes. This is because the information is encrypted.

D. Secure Computation Using Scalar Product

Secure scalar product techniques are used to privately compute record vector similarity. Yakout et al. [41] have come up with a technique for PPRL that involves two parties. The solution incorporates secure transformation in addition to SMC. As part of the preliminary protocol, string records are appended to vectors [47]. During this stage, the initial records are imbedded into a vector space by combining the fundamental reference sets with random string data. This is done in order to generate the initial vector. Each owner of the data supplies a vector of distances for each record, with each component of the vector containing the minimal Edit distance that exists between the record in question and the strings that comprise the i-th base set. It is the starting coefficient of the Discrete Fourier Transform for each vector, and it is represented as such. This map demonstrates that if two vectors are close to one another before mapping on the complex plane, then the values that are mapped from those vectors will be closed. This is demonstrated by the fact that the values that are mapped from those vectors are closed. As a result, the degree to which the initial data can be compared to one another is decided by the complex plane scalar product. In contrast to the secure transformation described in [47], the method described in [41] had both a high linkage utility (with no false negatives and only a minimal number of false positives) and a short execution time. Additionally, there were no false negatives and only a minimal number of false positives.

E. Hybrid Solution with Privacy-Preserving Blocking

In most cases, conducting an analysis of the pairwise similarity of all records is unnecessary. In point of fact, it is often sufficient to only look at attribute differences in order to identify the small portion of pairs that may be matching records and dismiss the remainder of the pairs. In order to make the most of this, the most recent solutions for PPRL make use of efficient blocking and indexing algorithms. These algorithms help to eliminate unnecessary pairwise comparisons while still keeping truly matched pairings. If the indexing/blocking strategy is not adequately designed, the blocking stage decreases matching overhead but may reduce utility. PPRL solutions usually follow the blocking step with an SMC phase to securely match data within the same block. Kuzu et al. came up with a three-party differentially private blocking method [48]. [49] This blocking method, which protects users' privacy, brings about a significant reduction in the cost of pairwise similarity evaluation at the SMC and produces reliable findings when it comes to linking datasets with personal identifiers.

F. PPRL Comparison

Table 1 compares the PPRL approaches from the preceding sections on five dimensions: a guarantee of user privacy, scalability, and the quality of linkages. First, in order to linkage, all PPRL systems require a third party, with the exception of Yakout et al. [41], which is the only exception. Second, indexing can be used to reduce the amount of computational work that is necessary, as shown by Yakout et al. [41] and Kuzu et al. [48] respectively. This was demonstrated in their respective studies.Last but not least, in terms of linkage quality, every solution except for Durham et al. [46] matches every record's attribute individually. Any one of the PPRL methods outlined above could potentially be utilised in a biological setting, depending on the circumstances. PPRL systems that make use of indexing techniques are more scalable for the linking jobs associated with Big Data. Third parties in protocols are also significant. The third party may simplify record matching, but it may also expose the protocol to collusion attacks.

Table 1: PPRL methods comparison

PPRL Methods	Privacy		Scalability		Linkage
	Protocols	#Parties	Indexing	Comp.	Matching
Durham [46]	Hashing Bloom Filter	3	None	Quadratic	Record
Yakout [41]	Embedding SMC	2	Sorted	Quadratic	Attribute
Scannapieco [47]	Embedding	3	None	Quadratic	Attribute
Kuzu [48]	DP & SMC	3	Spatial	Quadratic	Attribute

G. EHR Data Anonymization

Anonymization of EHR data safeguards sensitive private data and enables data analysis and study at the population level. It is challenging to create a dataset that is anonymous and that maximises the data's utility. This section examines the most up-to-date methods for sanitising individual records to ensure they comply with DP, working under the assumption that attackers have arbitrary prior knowledge. An early survey on methods of data publishing that respect users' privacy while also satisfying other privacy criteria is currently available. This survey will look at strategies. (which incorporates ideas such as k-anonymity, l-diversity, and t-closeness). We will apply DP algorithms to structured relational data, which is prevalent in biomedical data analysis, since most are designed for statistical databases. We cannot mention other intriguing unstructured clinical note research due to space constraints.

H. Partition-Based Methods

Partition-based approaches partition data and disturb each partition.

Dwork and colleagues [22] came up with a straightforward method by making use of individual Laplacian perturbations to be applied to the cell counts of the initial histogram. The addition of noise to the partition counts results in the creation of synthetic data. It does this by disturbing the elements of a nearly homogeneous partition together, which results in cost savings for the privacy budget. The differentially private spatial decompositions (PSDs) developed by Cormode et al. segment space into more manageable regions and provide statistics on the observations made in each segment's respective region. They provided two different strategies for partitioning utilising the quadtree and the data-dependent tree architectures (KD-tree). Only noises are necessary for the divider in order to safeguard the previous construction. In order to determine a private median for the latter structure, four different methodologies were suggested

The exponential technique is used in the division procedure in order to provide attributes by means of a

taxonomy tree using confidential data. DiffGen releases synthetic data after adding noise to leaf node counts. NoiseFirst and StructFirst, partition-based models by Xu et al. , differ in histogram structure computation and noise injection order. The first method obtains noisy cells [22] and merges adjacent cells in partitions using noisy counts. Second, choose borders using Exponential method and add Laplacian noise to the average of these divisions to create the optimum histogram. Due to computational complexity, these methods work best with low-dimensional data.

I. Transformation-Based Methods

Condensing the data into a compact representation (like bases, for example), and then manipulating that representation in order to generate synthetic data, is another approach.

The work done by Barak et al resulted in a two-step improvement being made to a differentially private Fourial perturbation algorithm (FPA). After computing a DP frequency matrix, they Fourier transformed it to the frequency domain by adding Laplacian noises to the Fourier coefficients. This brought the transformed matrix into the frequency domain. This brought the resulting matrix into the frequency domain. In order to produce sample fake data, a non-negative frequency matrix is reconstructed using linear programming. The fact that this model solves a linear programme with the same number of variables as the frequency matrix makes computation challenging. An enhanced discrete Fourier perturbation approach, also known as an EFPA, was developed by Ace et al. for differentially private histograms. They were able to improve FPA by removing high-frequency components and using the intrinsic connection of real-valued histogram Fourier coefficients. Additionally, they used a more exact score function for the Exponential process.

Jiang et al. developed a method of linear discriminant analysis that is based on principal component analysis. Before eigendecomposition, tweak mean and co-variance. They reconstructed a synthetic matrix using noisy eigenvectors. Too much noise in the co-variance matrix makes this technique unsuitable for high dimensions. Xiao and colleagues developed the Privelet method by first applying a wavelet transformation to the histogram (which is an invertible linear function), and then adding polylogarithmic noises to the mix. Privelet generates synthetic data by transformation, perturbation, and reconstruction, like PCA.

J. Statistical Model-Based Methods

These methods build statistical models from confidential data and then make sample points available to the public.

Machanavajjhala et al. built a differentially private data synthesiser by fitting private data to a multinomial dirichlet model and sampling from it. Typical parametric models have finite parameters. By employing a wide variety of sampling and filtering techniques, Cormode et al. were able to generate a compact histogram summary of sparse data that was subject to DP. Signals that are weaker than the cutoff threshold are weakened using a straightforward high-pass filter. This method randomly selects k cells from the contingency table that have count values of zero and then perturbs and releases every cell in the table that has a non-zero count. The values of these selected cells are then chosen at random from a particular distribution in order to match the output distribution of the baseline technique [22]. Two new methods featuring distinct priority sampling algorithms were developed as more sophisticated techniques.

The marginal distributions in this semi-parametric model are calculated non-parametrically, but the combined dependency of each dimension is represented parametrically by the correlation matrix. An additional method for the production of importance-weighted synthetic data was developed by Ji et al.

Using the methods described above, structured tabular data, such as demographics, can be anonymized. It is not possible to use previous methods to analyse these data because of the noise or the amount of computing required. Recent investigations shed light on these concerns, and we went over the relevant methodologies. They came up with a brand new algorithm for shrinking in order to enforce length limits.

K. Genomic Data Privacy

The cost of sequencing is going down, which means that high-throughput human genetic data may now be obtained at a lower cost for use in medical and biological research. Massive genomic data collection enables excellent diagnostic and treatment discovery. These potentials have led to several initiatives. Data based on genomes might also reveal information about other people. As a result, the potential danger to individuals' privacy may extend to their biological relatives [14]. Because of advances in both genetic research and hacking techniques, genomic data are now irreversible and present serious risks to individuals' privacy. The NIH has kept most aggregated results private due to privacy concerns [13]. Genome privacy has been protected using legal [2], ethical [15] [16], and technological [17] [19] means.

L. Secure Genomic Data Computation

Users are now able to store and analyse human genomic data by making use of cloud-computing services, thanks to the most recent modification to the rules governing the sharing of data by the National Institutes of Health (NIH). This alleviates concerns regarding the effective management of large genomic data sets. On the other hand, due to the fact that owners no longer have complete control over their data, the issue of privacy becomes a more pressing worry when cloud computing is used.

Using task-oriented optimizations (for example, a specialised data encoding approach), Lauter et al. [20] demonstrated that in order to obtain 80 bits of security, it is possible to do an analysis of 1000 genotype and phenotypic data in as little as 0.19 seconds to 6.85 seconds. Togan et al. [23] do research on the subject of HME-based comparisons of integers. HME is only capable of accurately computing small-scale edit distances for sequence lengths of less than 10. Kim et al. [21] demonstrated that HME is capable of providing an efficient and safe approximation of edit distance computation between two sequences of a length of 10,000 characters each. This was in reference to the challenges that are associated with scalability when utilising HEM. Moreover, Graepel et al. [24] and Naehrig [28] demonstrated that HME is able to make use of a variety of machine learning strategies. A approach for the homomorphic calculation of accurate logistic regression in

GWAS studies of rare illnesses has been presented by Wang et al. [37]. Zhang, Kim, and Lu developed an HME-based method for computing the chi-squared statistic.
Although HME-based secure outsourcing solutions for public clouds have the appearance of being promising, the reality is that HME is computationally intensive and calls for a significant amount of storage. There have been a lot of studies done on effective solutions that are task-specific. For instance, Ayday et al. [18] developed a method that protects patients' privacy and enables medical staff to access patients' short reads in a manner that is not visible to other staff members. Honey encryption is utilised by GenoGuard [25] to ward off brute-force assaults and preserve genetic data. [26] investigated methods of disease susceptibility testing that protected patients' privacy. Recent research published in [27] and [29] has proposed a wide variety of SMC-based safe genomic data analysis methods. A comparison of numerous different safeguarding strategies for the safe examination of genetic data was provided in Table 2.

M. Privacy-Preserving Genomic Data Dissemination

A generalised lattice method that is based on k-anonymity was proposed by Malin et al. [30] as a way to anonymize genomic sequences and ensure the security of genomic data transfer. Phenotypic-genotypic associations are safeguarded by the GWAS anonymization methodology developed by Loukides et al. [31]. A privacy-preserving logistic regression model employing DP was developed by Yu et al. [32] for the purpose of GWAS disease association identification. In the paper [33], Johnathon and colleagues developed a differentially private chi-square test statistic by leveraging genetic data. Uhler et al. [34] discovered an additional chi-squared test solution under differential privacy protection that releases the top M most important GWAS SNPs. This solution may be found in [link to article]. The Uhler methodology was improved upon by Yu et al. [37], [38] by incorporating improved utility and privacy tradeoffs as well as formal proofs. When creating high-dimensional genomic data, Zhao et al. [39] developed a method for the development of synthetic genomic data that makes use of linkage disequilibrium, which is a method for the reduction of features. This method helps to protect the privacy budget.

N. Privacy-Preserving Parallelization Techniques

It typically consists of two methods: the Map technique, which is used to organise data by mapping input key/value pairs to intermediate ones, and the Reduce method, which is used to summarise these intermediate pairs. Both of these methods are used for real-world activities. However, the computational paradigm utilised by the MapReduce architecture does not take into account any concerns regarding data security [40]. The provision of responsibility and access control for users, as well as the protection of privacy in computer models, are among the problems associated with security and privacy. Tran and Sato employed role-based access control (RBAC) and type enforcement in order to prevent malevolent MapReduce frameworks from leaking sensitive data. This was accomplished by enforcing specific data types (TE). Accountable MapReduce conducted an Accountability Test, also known as an A-Test, with the goal of locating rogue machines, which are also referred to as nodes that had optimised their resource utilisation in accordance with their allocated auditors. In the sense of protection methods, differential privacy (DP) and homomorphic encryption (HME) are both being researched for use with MapReduce in order to protect, respectively, the output and the process of aggregated computation. The researchers at Airavat coupled the MapReduce architecture with compulsory access control and differential privacy in order to ensure the safety and confidentiality of the outputs of aggregated processing. [21] Han et al. devised the DiffMR technique, which processed top-k searches in an exponential way, with the intention of preserving the differential privacy that had been established. Iterative selection that utilises reject rates that are established by the score function is carried out so that the accuracy of the query may be ensured even when it is applied to massive datasets. Chen et al. developed a privacy-protecting distributed technique for feature selection that is based on differential privacy. They used the MapReduce framework to apply Gini index-based approaches to large scale datasets. This allowed them to preserve users' privacy. Chen et al. utilised this particular algorithm. A heuristic approach to data anonymization was proposed by Zhang et al. in order to safeguard the privacy of individuals in an efficient manner. This approach finds intermediate datasets that are partially encrypted, as well as the link that leads back to their origin. When constrained optimization is carried out, it is done so with the intention of limiting the exposure of private information based on an upper bound on the jointly defined quantity of privacy that is lost across numerous datasets. This is done. Using a method called top-down specialisation (TDS), which is based on MapReduce, it is possible to anonymize massive datasets. In their paper, Zhang et al. developed a solution for hybrid cloud computing that would preserve users' privacy while simultaneously carrying out data-intensive computations. Sedic was able to modify MapReduce in such a way that it could partition the responsibilities connected with computing by sending data that had been sanitised to the public cloud and sensitive data to a private cloud. An Excalibur system was utilised by Santos et al. in their research in order to deliver policy-sealed interim data. Customers are responsible for deciding both the encryption and the decryption policies. Chen and Huang improved MapReduce so that it computes over encrypted intermediate data using fully homomorphic encryption. This was done by modifying MapReduce (FHE). PHE, which stands for parallel homomorphic encryption, is an encryption method that was designed with the intention of securely outsourcing the computation of massive datasets to a cluster of processors. On encrypted datasets, PHE made it possible to perform MapReduce operations such as element testing and keyword search, which allowed for the universal evaluation of parallelizable functions.
In the realm of genetic research, algorithms that preserve users' privacy have been developed for the goals of DNS read-mapping, association inquiry, and genomic signature search across huge datasets. These algorithms were designed for the objectives of using genomic research. The tamper-resistant characteristics of FPGAs were utilised to preserve intermediate data in MapReduce during the process of DNA read-mapping. Chen and colleagues came up with a solution that is able to perform hybrid cloud read mapping in a way that is both secure and scalable. Raisaro et al. came up with the idea for a parallelizable and flexible

privacy-preserving architecture for replication and fine-mapping genetic association studies spanning encrypted genotypes and phenotypes. This architecture would protect the confidentiality of the data. MapReduce was utilised to facilitate the parallelization of encrypted proteomic, transcriptomic, and metabolomic datasets. A Hadoop-based site-wise encryption approach for genomic signature discovery from entire human genome data was developed by Zhao et al.

III. CONCLUSION

Within the scope of biomedical research, we investigated a wide range of issues concerning the confidentiality of substantial amounts of data. The "big" component of data privacy is of the utmost importance since healthcare data usually comprise large scale clinical and genetic data. These data are huge in both size and dimension, making the "big" component of data privacy particularly significant. The reason for this is due to the fact that certain kinds of data are both huge in size and large in dimension. This makes it difficult to work with them. It may be difficult to find a solution to these one-of-a-kind problems given that they were not taken into account throughout the process of developing standard technology. For example, when it comes to dealing with whole genome sequencing (WGS) data, there are issues with the scalability of completely homomorphic encryption as well as secure multiparty computing solutions. These problems arise because of the complexity of the data involved. Other obstacles must be overcome in order to safeguard the outcomes of computations carried out on high-dimensional genomic data, and these obstacles have the potential to rapidly deplete the available resources in the event that careful planning is not applied to the process of disseminating the findings. We took a look at some of the most cutting-edge technologies currently available for record linking, synthetic data generation, and genomic data analysis, all of which are designed to safeguard individuals' privacy. We are of the opinion that a concerted effort from a wide variety of communities is necessary in order to find effective solutions to limit the privacy risks that are associated with biomedical research. Even though a lot of progress has been made, there are still a lot of problems and new obstacles that need to be handled. This is despite the fact that a lot of progress has been made. Despite the significant amount of progress that has been accomplished, this continues to remain the case (e.g, computer security, ELSI, biomedicine, genomics, etc.).

REFERENCES

[1] Health Information Technology for Economic and Clinical Health. 2010.

[2] L. Slaughter, Genetic Information Nondiscrimination Act of 2008, vol. 50. HeinOnline, 2008, p. 41.

[3] "Health Insurance Portability and Accountability Act (HIPAA)." [Online]. Available: http://www.hhs.gov/ocr/hipaa.

[4] D. Lafky, "The Safe Harbor method of de-identification: An empirical test," Fourth Natl. HIPAA Summit West, 2010.

[5] D. McGraw, "Why the HIPAA privacy rules would not adequately protect personal health records: Center for Democracy and Technology (CDT) brief," 2008. [Online]. Available: http://www.cdt.org/brief/why-hipaa-privacy-rules-would-not-adequately-protect-personal-health-records. [Accessed: 20-Sep-2015].

[6] K. Benitez and B. Malin, "Evaluating re-identification risks 103.with respect to the HIPAA privacy rule," J. Am. Med.Informatics Assoc., vol. 17, no. 2, pp. 169–177, 2010.

[7] P. Kwok, M. Davern, E. Hair, and D. Lafky, "Harder than you think: a case study of re-identification risk of HIPAA-compliant records," Chicago NORC Univ. Chicago. Abstr.,vol. 302255, 2011.

[8] L. Sweeney, "Data sharing under HIPAA: 12 years later," in Workshop on the HIP A A Privacy Rule's De-Identification Standard, 2010.

[9] S. J. Nass, L. A. Levit, and L. O. Gostin, Beyond the HIPAA Privacy Rule: Enhancing Privacy, Improving Health Through Research. The National Academies Press, 2009.

[10] X. Jiang, A. D. Sarwate and L. Ohno-Machado, "Privacy technology to support data sharing for comparative effectiveness research: A systematic review", Med. Care, vol. 51, no. 8 Suppl 3, pp. S58-565, Aug. 2013.

[11] B. A. Bernhardt, E. S. Tambor, G. Fraser, L. S. Wissow and G. Geller, "Parents' and children's attitudes toward the enrollment of minors in genetic susceptibility research: Implications for informed consent", Amer. J. Med. Genetics Part A, vol. 116, no. 4, pp. 315-323, 2003.

[12] A. L. McGuire et al., "To share or not to share: A randomized trial of consent for data sharing in genome research", Genetics Med., vol. 13, no. 11, pp. 948-955, 2011.

[13] N. Homer, S. Szelinger, M. Redman, D. Duggan, W. Tembe, J. Muehling, J. V Pearson, D. A. Stephan, S. F. Nelson, and D. W. Craig, "Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays.," PLoS Genet., vol. 4, no. 8, p. e1000167, Aug. 2008.

[14] M. Humbert, E. Ayday, J.-P. Hubaux, and A. Telenti, "Addressing the concerns of the lacks family: Quantification of kin genomic privacy," in Proceedings of the 2013 ACM SIGSAC conference on Computer & communications security, 2013, pp. 1141–1152.

[15] A. L. McGuire, T. Caulfield, and M. K. Cho, "Research ethics and the challenge of whole-genome sequencing.," Nat. Rev. Genet., vol. 9, no. 2, pp. 152–6, Feb. 2008.

[16] W. J. Dondorp and G. M. W. R. de Wert, "The 'thousand-dollar genome': an ethical exploration," Eur. J. Hum. Genet., vol. 21, pp. S6–S26, 2013.

[17] S. Sankararaman, G. Obozinski, M. I. Jordan, and E. Halperin, "Genomic privacy and limits of individual detection in a pool.," Nat. Genet., vol. 41, no. 9, pp. 965–7, Sep. 2009.

[18] E. Ayday, J. L. Raisaro, U. Hengartner, A. Molyneaux, and J.- P. Hubaux, "Privacy-Preserving Processing of Raw Genomic Data," Data Priv. Manag. Auton. Spontaneous Secur., vol. 8247, pp. 133–147, 2014.

[19] S. Wang, Y. Zhang, W. Dai, K. Lauter, M. Kim, Y. Tang, H. Xiong, and X. Jiang, "HEALER: homomorphic computation of ExAct Logistic rEgRession for secure rare disease variants analysis in GWAS.," Bioinformatics, vol. 32, no. 2, pp. 211–8, Jan. 2016.

[20] K. Lauter, A. López-Alt, and M. Naehrig, "Private computation on encrypted genomic data," in 14th Privacy Enhancing Technologies Symposium, Workshop on Genome Privacy . ht tp://seclab. soic. indiana.edu/GenomePrivacy/papers/Genome% 20Privacy-paper9. pdf.(29 July 2014, date last accessed), 2014.

[21] M. Kim and K. Lauter, "Private genome analysis through homomorphic encryption.," BMC Med. Inform. Decis. Mak., vol. 15 Suppl 5, no. Suppl 5, p. S3, Dec. 2015.

[22] C. Dwork, F. McSherry, K. Nissim and A. Smith, "Calibrating noise to sensitivity in private data analysis", Theory Cryptography, vol. 3876, no. 1, pp. 265-284, 2006.

[23] M. Togan and C. Plesca, "Comparison-based computations over fully homomorphic encrypted data," in

Communications (COMM), 2014 10th International Conference on, 2014, pp. 1–6.
[24] T. Graepel, K. Lauter, and M. Naehrig, “ML confidential: Machine learning on encrypted data,” in Information Security and Cryptology--ICISC 2012, Springer, 2013, pp. 1–21.
[25] Z. Huang, E. Ayday, J. Fellay, J.-P. Hubaux, and A. Juels, “GenoGuard: Protecting Genomic Data against Brute-Force Attacks,” in 36th IEEE Symposium on Security and Privacy, 2015.
[26] G. Danezis, “Simpler Protocols for Privacy-Preserving Disease Susceptibility Testing,” in 14th Privacy Enhancing Technologies Symposium, Workshop on Genome Privacy (GenoPri’14), 2014.
[27] Y. Zhang, M. Blanton, and G. Almashaqbeh, “Secure distributed genome analysis for GWAS and sequence comparison computation.,” BMC Med. Inform. Decis. Mak., vol. 15 Suppl 5, no. Suppl 5, p. S4, Dec. 2015.
[28] M. Naehrig, K. Lauter and V. Vaikuntanathan, "Can homomorphic encryption be practical?", Proc. 3rd ACM Workshop Cloud Comput. Secur. Workshop, 2011.
[29] S. D. Constable, Y. Tang, S. Wang, X. Jiang, and S. Chapin, “Privacy-preserving GWAS analysis on federated genomic datasets,” BMC Med. Inform. Decis. Mak., vol. 15, no. Suppl 5, p. S2, Dec. 2015.
[30] B. A. Malin, “Protecting genomic sequence anonymity with generalization lattices.,” Methods Inf. Med., vol. 44, no. 5, pp. 687–92, Jan. 2005.
[31] G. Loukides, A. Gkoulalas-Divanis, and B. Malin, “Anonymization of electronic medical records for validating genome-wide association studies,” Proc. Natl. Acad. Sci. U. S. A., vol. 107, no. 17, pp. 7898–7903, 2010.
[32] F. Yu, M. Rybar, C. Uhler, and S. E. Fienberg, “Differentially- Private Logistic Regression for Detecting Multiple-SNP Association in GWAS Databases,” in Privacy in Statistical Databases, vol. 8744, J. Domingo-Ferrer, Ed. Cham: Springer International Publishing, 2010, pp. 170–184.
[33] A. Johnson and V. Shmatikov, “Privacy-preserving data exploration in genome-wide association studies,” in Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD ’13, 2013, p. 1079.
[34] C. Uhler, A. B. Slavkovic, and S. E. Fienberg, “Privacy-preserving data sharing for genome-wide association studies,” J. Priv . Confidentiality , vol. 5, no. 1, pp. 137–166, 2013.
[35] A. K. Elmagarmid, P. G. Ipeirotis and V. S. Verykios, "Duplicate record detection: A survey", IEEE Trans. Knowl. Data Eng., vol. 19, no. 1, pp. 1-16, Jan. 2007.
[36] [36] L. Getoor and A. Machanavajjhala, "Entity resolution: Theory practice & open challenges", Proc. VLDB Endow., vol. 5, no. 12, pp. 2018-2019, Aug. 2012.
[37] F. Yu, S. E. Fienberg, A. B. Slavković, and C. Uhler, “Scalable privacy-preserving data sharing methodology for genome- wide association studies.,” J. Biomed. Inform., vol. 50, no. 50C, pp. 133–141, Feb. 2014.
[38] F. Yu and Z. Ji, “Scalable Privacy-Preserving Data Sharing Methodology for Genome-Wide Association Studies: An Application to iDASH Healthcare Privacy Protection Challenge,” BMC Med. Informatics Decis. Mak. [submitted], 2014.
[39] Y. Zhao, X. Wang, X. Jiang, L. Ohno-Machado, and H. Tang, “Choosing blindly but wisely: differentially private solicitation of DNA datasets for disease marker discovery.,” J. Am. Med. Inform. Assoc., vol. 22, no. 1, pp. 100–8, Jan. 2015.
[40] D. Chen and H. Zhao, “Data Security and Privacy Protection Issues in Cloud Computing,” 2012 Int. Conf. Comput. Sci. Electron. Eng., vol. 1, no. 973, pp. 647–651, 2012.
[41] M. Yakout, M. J. Atallah and A. Elmagarmid, "Efficient and practical approach for private record linkage", J. Data Inf. Quality, vol. 3, no. 3, pp. 5:1–5:28, Aug. 2012.
[42] A. Al-Lawati, D. Lee and P. McDaniel, "Blocking-aware private record linkage", Proc. 2nd Int. Workshop Inf. Quality Inf. Syst., pp. 59-68, 2005.
[43] H.-C. Kum, A. Krishnamurthy, A. Machanavajjhala, M. K. Reiter and S. C. Ahalt, "Privacy preserving interactive record linkage (PPIRL)", J. Amer. Med. Inf. Assoc., vol. 21, no. 2, pp. 212-220, 2014.
[44] B. H. Bloom, "Space/time trade-offs in hash coding with allowable errors", Commun. ACM, vol. 13, no. 7, pp. 422-426, Jul. 1970.
[45] M. Kuzu, M. Kantarcioglu, E. Durham and B. Malin, "A constraint satisfaction cryptanalysis of bloom filters in private record linkage", Proc. 11th Int. Symp. Privacy Enhancing Technol., vol. 6794, pp. 226-245, 2011.
[46] E. Durham et al., "Composite Bloom filters for secure record linkage", IEEE Trans. Knowl. Data Eng., vol. 26, no. 12, pp. 2956-2968, Dec. 2014.
[47] M. Scannapieco, I. Figotin, E. Bertino and A. K. Elmagarmid, "Privacy preserving schema and data matching", Proc. 2007 ACM SIGMOD Int. Conf. Manage. Data, pp. 653-664, 2007.
[48] M. Kuzu, M. Kantarcioglu, A. Inan, E. Bertino, E. Durham and B. Malin, "Efficient privacy-aware record integration", Proc. 16th Int. Conf. Extending Database Technol., pp. 167-178