Published in final edited form as:
Ann N Y Acad Sci. 2017 January ; 1387(1): 73–83. doi:10.1111/nyas.13259.

Genome privacy: challenges, technical approaches to mitigate risk, and ethical considerations in the United States

Shuang Wang[1,a], **Xiaoqian Jiang**[1,a], **Siddharth Singh**[1], **Rebecca Marmor**[1], **Luca Bonomi**[1], **Dov Fox**[2], **Michelle Dow**[1], and **Lucila Ohno-Machado**[1]

[1]Department of Biomedical Informatics, University of California San Diego, La Jolla, California

[2]School of Law, University of San Diego, San Diego, California

Abstract

Accessing and integrating human genomic data with phenotypes is important for biomedical research. Making genomic data accessible for research purposes, however, must be handled carefully to avoid leakage of sensitive individual information to unauthorized parties and improper use of data. In this article, we focus on data sharing within the scope of data accessibility for research. Current common practices to gain biomedical data access are strictly rule based, without a clear and quantitative measurement of the risk of privacy breaches. In addition, several types of studies require privacy-preserving linkage of genotype and phenotype information across different locations (e.g., genotypes stored in a sequencing facility and phenotypes stored in an electronic health record) to accelerate discoveries. The computer science community has developed a spectrum of techniques for data privacy and confidentiality protection, many of which have yet to be tested on real-world problems. In this article, we discuss clinical, technical, and ethical aspects of genome data privacy and confidentiality in the United States, as well as potential solutions for privacy-preserving genotype–phenotype linkage in biomedical research.

Keywords

genome privacy and confidentiality; record linkage; ethics; privacy law

Introduction

With the proliferation of genome sequencing technologies, human genomic data have been widely used in biomedical research studies. The availability of such data, together with the enhanced capacity to process them, is leading to advancements in biomedical science, informatics, and bioethics. One example is genome-wide association studies (GWASs) that

Address for correspondence: Shuang Wang, Health System Department of Biomedical Informatics, University of California San Diego, 9500 Gilman Dr., MC 0728, La Jolla, CA 92093. shw070@ucsd.edu.

[a]Contributed equally to this work.

S.W. and X.J. wrote major parts of this manuscript; S.S. and R.M. made contributions to the section on clinical environments; LB contributed to the patient linkage section; D.F. contributed to the legal and ethical components; M.D. helped edit the text and figures; and L.O.M. provided the motivation and detailed edits to the manuscript.

Conflicts of interest
The authors declare no conflicts of interest.

attempt to identify single-nucleotide variants or polymorphisms (SNPs) associated with a disease/phenotype of interest. Researchers can benefit from the convenient access to aggregate human DNA data, in particular the allele frequencies of thousands to millions of SNPs across human populations with certain diseases or medical conditions, but data privacy and confidentiality concerns need to be addressed. Data privacy and confidentiality are relevant, but there are differences that should be emphasized. Data confidentiality focuses on keeping the data secure and private from unauthorized access, ensuring the data fidelity in storage or during transferring. Data privacy is concerned with the appropriate use of data. That is, data should be used according to intended purposes without violating patient intentions. Data privacy is certainly tied to data confidentiality. In general, strong data privacy is not possible without having an appropriate data confidentiality protection program. But data confidentiality practices may not always guarantee data privacy. For example, data privacy may be compromised if an authorized user of a "de-identified" data set attempts to re-identify patients or infer information that could compromise patient privacy.

Open-access mechanisms (e.g., as in the 1000 Genomes Project[1] and the Personal Genome Project, PGP[2]) and controlled-access mechanisms (e.g., as in the database of Genotypes and Phenotypes, dbGaP[3]) are widely used for genomic data sharing. Compared to controlled-access genomic datasets (e.g., dbGaP), open-access datasets have been used in many more studies in a given year.[4] Broad genomic data sharing through an open-access model can benefit the public in advancing scientific discovery; however, such models also create new re-identification risks, as demonstrated by recent studies.[5,6] Although such re-identification risks may not be applicable to participants who are interested in sharing their genomic data and other information through public platforms (e.g., DNA.Land or Open Humans), we focus here on unauthorized re-identification. For example, a study conducted by Gymrek *et al.*[5] reported that about 50 individuals could be re-identified (i.e., participants' names) in the 1000 Genomes Project by leveraging an online genealogy database. In response to this risk, the National Institutes of Health (NIH) removed the age information of participants in this project.[4] The re-identification risk of the PGP was also reported by Sweeney *et al.*[6] In an analysis by Homer *et al.*[7] of the attribute disclosure risk using data from dbGaP, it was found that allele frequencies of an individual can be used to reliably determine his/her presence in a case group. As a result, in addition to widely discussed privacy risks (e.g., authorized re-identification, forensic or other law enforcement re-identification[8,9]), participants of human genome studies (HGSs) may also be subject to attribute disclosure risk even from what are labeled "de-identified data."[10] In response to this threat, the NIH shifted to a controlled-access model, which removes all aggregate data from open-access repositories to protect HGS participants against re-identification.[11]

Today, HGSs are most frequently shared in a controlled-access manner (e.g., as in dbGaP[3]) that prevents broad dissemination of data. There has been intensive debate about such a controlled-access model, as the approval process may prevent researchers from timely gaining data access. As discussed by Zhou *et al.*,[12] "some researchers pointed out that the NIH may have overreacted, as the attack power achievable over at least some data-sets can be very limited. On the other hand, such agreement-based protection has been found to be insufficient, as confidential user information can still be derived from other public sources."

As an example, a previous study[13] showed that test statistics, such as *P* values, in GWASs could disclose a significant amount of personal information, such as identifying individuals or inferring portions of their genomic data.

The biomedical community recognizes that, for large-scale genomics projects, absolute privacy and data protection cannot be ensured in practice and should not be promised to participants.[14,15] But rapid dissemination of research findings and related data are critical for scientific and technological progress in HGSs and biomedicine. Therefore, attempts have been made to share genomic data (i.e., allele frequencies of genomic variants) in a way that the privacy risk and the benefit of timely sharing of the data can be balanced. An example is Geno2MP, a web-based tool that allows the query of rare variants from sequencing data linked to one or multiple Mendelian phenotypes defined by human phenotype ontology (HPO) terms.[16] As another example, the Cancer Genome Atlas refrains from releasing non-validated somatic mutations inferred from cancer sequencing data,[17] but the actual privacy risk in these data has never been systematically evaluated and thus likely varies from one individual to another. Such efforts, however, are mostly ad hoc, lacking rigorous theoretical foundations for setting the boundary where the balance should be struck. Figure 1 summarizes the existing problems related to genome privacy in biomedical research.

Another important privacy challenge comes from the requirement of private data exchange in cross-institutional studies. For example, data from the same individual may be partitioned among several sites, such as healthcare providers, sequencing facilities, insurance companies, and research institutions. There are incentives to study them jointly. For example, the genomic data of a particular group hosted in a sequencing facility can be significantly enriched by linking such data to electronic health records (EHRs),[18] allowing comprehensive and simultaneous capture of multiple exposures, health status, interventions, and outcomes. In this case, record linkage is an essential first step to combine data in cross-institutional studies.[19] For example, Bozkurt *et al.*[20] demonstrated that linking a pharmaceutical database to a biobank database can help in the discovery of interactions between thiazide diuretics and genetic variation for type 2 diabetes. In such cross-institutional studies, data privacy and confidentiality concerns are prominent problems, as the study may require patient-level data exchange for record linkage. The exposure of protected health information (PHI) can put individuals' sensitive information at risk.

Despite several recent record linkage efforts, there are many challenges in ensuring accuracy and privacy in record linkage. For example, false positives may result from the incorrect linkage of information of two different individuals in vertically partitioned datasets, which can result in estimation errors in research studies. Vertically partitioned datasets are those in which part of the information related to an individual is stored in one dataset, and another part is stored in a separate dataset. Existing record linkage methods can be categorized into deterministic[21] and probabilistic[22] approaches. If there are explicit identifiers (e.g., name, social security number) among different datasets, deterministic record linkage methods are often adopted. Probabilistic linkage methods are more complex, as they need to assign different weights for different discriminative linkage variables to compute an overall score that indicates how likely a record pair may come from the same individual. Furthermore, owing to privacy and security concerns, institutions and individuals may be hesitant to share sensitive personal health information outside the health system. Hence, robust privacy-preserving record linkage tools are needed before this rich environment is ripe for research use.

In this paper, we focus on genomic data privacy from clinical, technical, and ethical perspectives, as well as technologies that facilitate privacy-preserving genomic data sharing and integration with phenotypes through privacy-preserving record linkage.

Genome privacy and privacy-preserving record linkage in clinical environments

With the Health Information Technology for Economic and Clinical Health (HITECH) Act, the U.S. government mandated the implementation of EHRs to improve quality of care, and by 2013, 59% of hospitals had adopted EHRs.[23] This has opened doors for creating large-scale, accurately phenotyped electronic cohorts of individuals with specific conditions, such as inflammatory bowel disease[24] and rheumatoid arthritis,[25] by combining structured and narrative data from these EHRs with natural language processing tools. At the same time, the decreasing costs and rapid sequencing techniques have enabled large-scale HGSs. To promote meaningful discovery research through a combination of EHRs and genomic data, consortia such as the Electronic Medical Records and Genomics (eMERGE) Network, funded by the National Human Genome Research Institute, have emerged with the intention of studying new methods and tools to use better EHR data in genomic research.[26] The eMERGE Network includes geographically different groups, where each group has its own biorepository. Phenotypic data within EHRs can be linked to each group's genomic data derived from its biorepository, and these data can be aggregated from different institutions, by which a large number of phenotypes for both case and control patients can be collected in an efficient manner for discovering genotype–phenotype associations. While this is a very promising approach with tremendous potential for advancing science and clinical care, it is still not widespread across a large number of institutions and a large number of health conditions, as inadequate genome privacy preservation can quickly result in a loss of public trust.

Technical solutions for privacy-protecting disclosure of data or results from genome studies

Existing technical solutions

In this section, we focus on the technical solutions for genome privacy and categorize the discussion into three parts as (1) privacy-preserving aggregate statistics/data disclosure, (2) data outsourcing in an untrusted cloud,[27] and (3) privacy-preserving cross-institutional collaboration. There are several studies that have focused on the protection of genome research outcomes[28–31] and genomic data dissemination[32] using differential privacy (DP),[33] which aims to protect the privacy of an individual from being breached, when the individual has opted in or out of a study. The protect strength is controlled by a privacy budget, where a smaller budget provides a stronger protection, and vice versa. Perturbation-based methods are widely used to achieve differential privacy, where the perturbation noise is calibrated on the basis of the privacy budget and sensitivity (i.e., the maximum degree of value change for a function of interest, given the absence/presence of an individual). The Laplace mechanism[33] and the Exponential mechanism[34] are two commonly used methods to achieve DP. More specifically, the Laplace mechanism calculates the DP outputs by adding noise, which consists of randomly sampled values from the Laplace distribution. In contrast, the Exponential mechanism generates DP outputs by drawing random samples based on a user-defined utility function. For the technical details of DP, the reader is referred to other published articles.[33,34] The Laplace mechanism was employed to add noise to the chi-squared statistics to protect the release of most signification SNPs identified in GWASs.[28] Both the Laplace and Exponential mechanisms were used in order to protect the dissemination of medical and genomic data.[32,35] Differential privacy-based methods provide strong and provable protection on genomic data privacy. However, these perturbation-based protection methods often introduce too much noise and render some results untrustworthy for practical genomic applications.

In contrast, many studies have been conducted to safeguard genomic data analysis tasks in secure outsourcing[36] using Homomorphic Encryption (HME),[37] which allows users to directly perform certain arithmetic operations over encrypted data. There are different versions of HME:[38] (1) partial HME only supports computations that can be expressed as either addition or multiplication operations over ciphertext;[39] (2) full HME supports an unlimited number of addition and multiplication operations, but often has formidable computation costs;[40] and (3) quasi HME is specified by a certain number of accumulated multiplication operations[41] and is most flexible and efficient for a specific task. Studies that use HME mainly focus on building secure primitives using the combination of HME addition and multiplication operations to achieve accurate and efficient computation of certain genome analysis tasks (e.g., chi-squared statistics computation,[42,43] Hamming and edit distance comparison,[44,45] regression model learning and evolution[36]).

For privacy-preserving collaboration, garbled circuit[46,47] and secret sharing[48,49] schemes have been applied to enhance genome privacy protection in a federated computational environment. A garbled circuit (also known as Yao’s protocol[50]) can transform any function into a secure circuit representation for secure computation between two parties in a “semi-honest” model, in which each party follows the protocol exactly but may be curious about the other party’s data. Secret sharing[51] is another approach based on an encrypted Boolean circuit that can support secure computation over more than two parties in a semi-honest model, where encrypted data are securely distributed among all participating parties. All parties can collaborate to securely evaluate a Boolean circuit for a certain task. Unlike many existing studies[52–56] that protect individual-level data, these techniques based on secure multiparty computation can also protect the exchange of intermediary information during the entire computation process. Although it is appealing to apply a cryptographic method for secure outsourcing and privacy-preserving collaboration, only limited genome computations[47,48,57] have been supported to date. Further investigation of these types of privacy-protection technologies is still necessary.

Community efforts for genome privacy

As a first step toward systematic protection of clinical genomic data and to enable their convenient and privacy-preserving dissemination, we organized the Critical Assessment of Data Privacy and Protection (CADPP) challenges at the iDASH National Center for Biomedical Computing Privacy Workshop to solicit community help to better understand and mitigate privacy risks in genomic data dissemination[58] and privacy-preserving analysis.[59] The competition was attended by leading data privacy groups, together with biomedical, genetics, and bioethical experts, and was reported by GenomeWeb[60] and Nature News.[61] We evaluated the utility based on the probabilities that the highly significant SNPs can be preserved after adding noise.[58] For example, the 84.8% true significant top 10 SNPs can be preserved on a small dataset with 5000 SNPs based on differential privacy protection. We also evaluated the privacy risk of releasing a DP genomic dataset on the basis of the likelihood ratio (LR) test.[62] Sankararaman *et al.*[62] showed that the LR test is one of the strongest re-identification tests and that it can provide the upper bound of the re-identification power with the independent SNP assumption. The experimental results showed that the best DP-based protection algorithm over a small dataset of 311 SNPs can achieve a re-identification power as low as 0.01, which indicates the ratio of re-identifiable participants in a case group at the statistical power of 0.95 in GWASs. The outcome of the competition revealed practical challenges in protecting high-dimensional genomic data, even when the data were aggregated from a large number of individuals (e.g., allele frequencies): once the number of SNPs to be protected exceeds a few hundred, it becomes difficult to share allele data directly without exposing the identities of some individuals or destroying the utility of the data through added noise. Computational and communication overheads are still significant for cryptographic technologies used in genome privacy protection. On a positive note, we found that privacy-preserving techniques work well when the results of a whole GWAS (instead of raw allele frequencies) are shared: even when holding results to a high privacy standard (i.e., using differential privacy[33]), most utility can still be conserved if a small number (e.g., 5 to 10) of highly significant SNPs are made public. In addition, by introducing approximations in the edit distance computation based on the characteristics of the human genome, cryptographic technologies can be efficiently optimized to handle large-scale, privacy-preserving human genomic data analyses.

Recent advances in algorithm development allow computation of multivariate models over vertically partitioned datasets and may be particularly helpful when genome data are hosted in one institution and EHRs are hosted in another for study participants.[43] Computing over vertically partitioned data decreases the need for moving data around and thus helps protect individual records, but this type of distributed computation can only be achieved if records are linked across vertical partitions. Record linkage is also needed in cases where researchers want to enrich existing study data (e.g., dbGaP records) with longitudinal outcomes prospectively collected from EHRs.

Technical solutions for record linkage

In this section, we discuss existing technical solutions for record linkage. More specifically, we cover deterministic and probabilistic methods as well as privacy-protection techniques to mitigate loss of privacy risks in record linkage. Figure 2 depicts examples of record-linkage systems for vertically partitioned data split into a hospital and an external biobank (where DNA sequencing data are available).

Deterministic record linkage

Deterministic methods determine whether record pairs represent a linkage based on a fixed criterion (i.e., matching rule) involving a set of specific identifiers/variables, which can contain either original individual identifiers[63,64] (e.g., full name, birthday, full address) or partial information on the original individual identifiers[65,66] (e.g., first four letters of surname, birth year, partial address). For example, record pairs can be matched by comparing the name, date of birth, gender, and zip code. The matching result can be assessed by two different methods depending on the type of comparison between the records variable values. Existing deterministic methods can be categorized into exact matching[63–66] and approximate matching.[21] In exact matching, a record pair is considered a match if and only if all the linkage-variable values are exactly the same across the records. On the other hand, in approximate matching,[21] linkage variables are compared in a less strict way, which takes into consideration the similarity between the variable values. For example, in matching values referring to names, string similarity measures are typically employed to measure the closeness between the names. As a result, records with minor dissimilarity in their string values (e.g., typos, abbreviated names) can be still identified as matching records. Therefore, in the presence of data heterogeneity, approximate record-linkage solutions are preferred over the exact-matching approaches.

Probabilistic record linkage

In contrast to deterministic solutions, probabilistic methods employ a statistical approach to determine whether record pairs represent a match. In probabilistic linkage methods, a wider range of linkage variables is taken into consideration and algorithms assign different weights to different linkage variables on the basis of their discriminatory power (i.e., frequency, uniqueness of data) and the possible presence of errors.[22] The main purpose of variable weights is to statistically model the ability of each individual variable to correctly identify a match or a non-match. For each pair of records, a weighted composite score is computed, representing the probability of the two given records belonging to the same entity. Specifically, a decision is made by comparing the composite score against two threshold values, which serve as cutoff lines to determine the matching and non-matching pairs, and those that may represent possible matching records but require manual revision (i.e., human intervention). While the computation of best thresholds is an open question, depending on the final application the appropriate values can be determined in different ways (e.g., by minimizing the error probability of making an incorrect decision, based on previous experiments[67] or manual inspection).[68] Owing to the use of statistical and data analysis methods that compute and aggregate the variable weights, probabilistic record-linkage approaches yield higher utility when compared to deterministic methods and tend to be more robust to noise and missing values.

Privacy-preserving record linkage

Record linkage usually requires the exchange of sensitive individual identifiers as linkage variables, which raises significant privacy concerns. Consequently, many efforts have been made to protect the exchanged sensitive information in individual record linkage.[69–72] Among existing solutions, secure transformation (e.g., one-way hashing),[72,73] hybrid solutions,[74,75] and cryptography-based solutions[76] have been developed.

Secure transformation methods link records based on transformed information of the original data (e.g., names, address), from which sensitive information cannot be easily recovered. The typical scenario requires a trusted third party. First, all data owners apply the same secure transformation scheme to generate a new representation of their private data. Second, the transformed information is sent to the third party to perform comparison. Many studies have been conducted to improve privacy protection in the transformation phase (e.g., one-way hashing,[77] attribute generalization,[75] n-grams,[70] embedding,[71] cryptography[78]). For example, Kho *et al.*[77] developed a hash-based privacy-protecting record-linkage system and evaluated it across six institutions in Chicago, covering more than 7 million records. To rank the matching similarity, the authors designed 17 seeded hash code combinations of individual identifiers (e.g., name, birthday).[77] Although the underlying sensitive individual identifiers among different parties[77] were concealed by the one-way hashing, there was still a risk of information disclosure. For instance, given the hash function, a curious user from institution A could hash a dictionary of names to check the existence of certain individuals from other institutions. To address these concerns, it is essential to include a trusted authority to perform hash comparison in one-way hashing–based linkage methods. However, a trusted authority may not always exist for ad hoc data analysis collaboration, which limits the usability of the one-way hashing–based linkage method.

N-grams (i.e., substrings of length *n* derived from the original individual's identifiers) are used in another data transformation method for privacy-protecting linkage methods.[79] However, since, in the N-gram algorithm, substrings of original linkage variables may still leak partial information of individuals' sensitive information, hybrid methods have been studied to combine both anonymization and transformation techniques. For example, a DP algorithm was developed to protect individual information by perturbing the frequencies of N-grams.[71] Unfortunately, noise injected by the DP method may result in degraded linkage accuracy and therefore represent tradeoffs of privacy protection and linkage accuracy.

Adam *et al.*[78] proposed a cryptography-based solution for privacy-preserving integration of healthcare data from different sources. The key idea of their approach is to use commutative encryption to encrypt all sensitive data by all of the data sources, using their own keys. By using commutative encryption algorithm, data from different sources can be encrypted sequentially with different keys from different parties, where the output ciphertext does not depend on the order of the encryption. Therefore, the encrypted value across different datasets will be identical if and only if their plaintexts are the same. Then, the proposed technique is able to compare encrypted data across different sources.

Ethical and legal implications

The encryption of genomic data for storage and computation raises a number of important ethical and legal implications related to informed consent and privacy. Genomic studies typically seek to protect research subjects' personal information by removing potentially identifying information apart from potentially including coded identifiers. Such de-identified information is exempt from privacy and other protections under the federal regulations that govern research on human subjects more generally.

Therefore, research subjects have limited legal protection for their de-identified genomic information. As we explain later, existing privacy laws and regulations exempt de-identified data obtained from research subjects, since personal identifiers are removed. However, it might still be possible to re-identify individuals through a range of methods, such as side-channel leaks.[80] These strategies use computerized network databases that cross-reference these data against sources such as voter registration databases and other records that can contain identifying information, such as a person's gender, surname, ZIP code, and date of birth.[6]

Informed consent

The risk of re-identifiying patient health information presents problems for informed consent systems. Most federally funded human subjects research must abide by the informed consent requirements of the Common Rule.[81] Research that uses data from which it is difficult—but not impossible—to determine the identity of individual subjects, however, is exempted from requirements to advise and consent subjects about relevant risks, including privacy risks. As long as personal identifiers have been removed from the genetic information provided by subjects, researchers need not advise or obtain consent from participants about any new privacy risks—however serious—about the risk of their information being re-identified.[82]

De-identified data that are stored in large-scale genomic biobanks are often obtained using cursory informed consent forms, sometimes referred to as broad consent.[83] Research studies on those data tend to use broad consent forms as well.[83] But consent is not meaningful unless it is obtained in a way that actually informs research subjects about the relevant risks of participation. When the risks are instead communicated in an excessively generic or ambiguous way, the process may leave subjects uninformed about the risks that are possible if they participate. An example of this kind of broad consent is when subjects are told that their data "may be used for a range of projects by researchers worldwide."[84] Such open-ended communications are unlikely to inform research subjects[85] and therefore unlikely to be considered legally valid. This superficial disclosure comes at the beginning of the research process and implicates the use of personal material and information for years thereafter, without indicating the specific lines of research to be conducted using the subjects' samples. New consent forms need not be secured for these unspecified studies over untold duration unless state laws require it.[86]

Study participants have the right to withdraw consent to the use of their data. Specifically, they can withdraw their previously given consent to use their data until the point at which those data have been analyzed or aggregated for publication.[87] But data-sharing and deidentification procedures make it difficult for research subjects to withdraw their data once consent has been obtained.[88] There is no immediate way to retract data or restore privacy in the event of a breach. Withdrawal is especially difficult in international studies, as broad sharing of samples and specific regulations may significantly complicate or disallow the process of tracing and removal of individually derived data.[89]

Privacy

Genomic data for which individual identifiers have been removed are unprotected under the major U.S. privacy law, namely, the Health Insurance Portability and Accountability Act (HIPAA).[90] Research using these data is considered not to involve human subjects; therefore, informed consent does not need to be obtained and no special protection for the data is mandated, because HIPAA's privacy protections do not apply to such de-identified data. The Genetic Information Nondiscrimination Act (GINA),[91] while restricting the collection, use, and disclosure of personal data by certain entities, does not specifically regulate genomic data access. Instead, GINA bars large employers from discriminating against employees on the basis of their genetic test results or family histories. It also creates a legal basis for people to seek legal action against health insurance companies in case their genetic information is used to determine premiums or coverage.

Conclusion

Very limited legally protected interests in personally sensitive, imperfectly anonymous data makes cryptographically secure protocols for genomic data absolutely critical. These solutions seek to make data usable for important purposes of medical research and scientific understanding, while effectively preventing unauthorized operations. However, most cryptographic protocols are beset by common weaknesses, as they are applied to GWASs, such as scalability issues in handling real-world large-scale genomic data, flexibility concerns in dealing with complicated analysis tasks, reliance on semi-trusted entitles, and degradation of data utility due to noise injection methods. Cryptography solutions have yet to resolve privacy-preserving data sharing for genomic association. Human genomic data are becoming increasingly important to biomedical research but they cannot be easily shared owing to their sensitive nature. In this article, we discussed both technical and ethical accepts of genome privacy in the context of biomedical research. It is important to develop tools to help biomedical researchers share, evaluate, or choose the most useful genomic data for research in a privacy-protecting manner, without undermining the utility of the data. Because genomic data are often studied together with other information, we also covered record linkage, an essential step to make use of partitioned information across institutions. We summarized existing solutions for privacy-preserving record linkage in healthcare settings and identified several challenges, including, but not limited to, linkage accuracy and privacy concerns in exchanging sensitive linkage data.

The development of a well-balanced policy for genome and clinical data sharing requires joint efforts from technical, regulatory, and ethics communities, including experts in computer science, computer security, genetics, ethics, privacy law, and many other fields, to enable efficient genomic data analysis and record linkage in biomedical research, while respecting the privacy of individuals.

We have limited our focus in this article to the United States, with its patchwork of legal regulations on data privacy. U.S. privacy protections are scattered across different laws, generally divided by data content, potential uses, and funding sources (e.g., Children's Online Privacy Protection Act (COPPA), GINA, HIPAA). Distinct challenges arise in the European Union (E.U.), which has an overarching legal framework (i.e., EU Data Protection Directive) for all types of data privacy.[92] Article 8 of the Charter of Fundamental Rights of the European Union recognizes data protection as a separate fundamental right (even as there remains variation in how E.U. member states implement these rules).[93] Important differences between the United States and the European Union, as well as other regions, warrant further investigation along this line.

Acknowledgments

This work was funded in part by National Institutes of Health Grants U54HL108460, R00HG008175, UL1TR001442, T15LM011271, R00LM011392, and R21LM012060, and Patient-centered Outcomes Research Institute (PCORI) Contract CDRN-1306-04819.

References

1. 1000 Genomes Project. Retrieved April 10, 2016 from http://www.1000genomes.org/
2. Church GM. The Personal Genome Project. Molecular systems biology. 2005; 1
3. Mailman, Matthew D., Feolo, Michael, Jin, Yumi, et al. The NCBI dbGaP database of genotypes and phenotypes. Nature genetics. 2007; 39(10):1181–1186. [PubMed: 17898773]
4. Rodriguez, Laura L., Brooks, Lisa D., Greenberg, Judith H., Green, Eric D. The complexities of genomic identifiability. Science. 2013; 339(6117):275–276. [PubMed: 23329035]
5. Gymrek, Melissa, McGuire, Amy L., Golan, David, Halperin, Eran, Erlich, Yaniv. Identifying personal genomes by surname inference. Science. 2013; 339(6117):321–324. [PubMed: 23329047]
6. Sweeney, Latanya, Abu, Akua, Winn, Julia. Identifying Participants in the Personal Genome Project by Name (A Re-identification Experiment). 2013.
7. Homer, Nils, Szelinger, Szabolcs, Redman, Margot, et al. Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. PLoS genetics. 2008; 4(8):e1000167. [PubMed: 18769715]
8. Beskow, Laura M., Dame, Lauren, Costello, E Jane. Research ethics. Certificates of confidentiality and compelled disclosure of data. Science (New York, NY). 2008; 322(5904):1054–1055.
9. Kaye, Jane. Police collection and access to DNA samples. Genomics, Society and policy. 2006; 2(1):16–27.
10. Couzin, Jennifer. Genetic privacy. Whole-genome data not anonymous, challenging assumptions. Science (New York, NY). 2008; 321(5894):1278.
11. Zerhouni, Elias A., Nabel, Elizabeth G. Protecting aggregate genomic data. Science (New York, NY). 2008; 322(5898):44.
12. Zhou, Xiaoyong, Peng, Bo, Li, Yong Fuga, Chen, Yangyi, Tang, Haixu, Wang, XiaoFeng. Computer Security--ESORICS 2011. Springer-Verlag; 2011. To release or not to release: evaluating information leaks in aggregate human-genome data; p. 607-627.
13. Wang, Rui, Li, Yong Fuga, Wang, XiaoFeng, Tang, Haixu, Zhou, Xiaoyong. Proceedings of the 16th ACM conference on Computer and communications security - CCS ’09. ACM Press; 2009. Learning your identity and disease from research papers; p. 534-544.
14. Lunshof, Jeantine E., Chadwick, Ruth, Vorhaus, Daniel B., Church, George M. From genetic privacy to open consent. Nature Reviews Genetics. 2008; 9(5):406–411.

15. Prainsack, Barbara, Buyx, Alena. A solidarity-based approach to the governance of research biobanks. Medical Law Review. 2013; 21(1):71–91. [PubMed: 23325780]
16. Geno2MP. Retrieved September 18, 2015 from http://geno2mp.gs.washington.edu/Geno2MP/#/
17. National Human Genome Research Institute. The Cancer Genome Atlas (TCGA). Retrieved June 10, 2014 from https://tcga-data.nci.nih.gov/tcga/
18. Pukkala, Eero. Methods in Biobanking. Springer; 2011. Biobanks and registers in epidemiologic research on cancer; p. 127-164.
19. Christen, Peter. Data matching: concepts and techniques for record linkage, entity resolution, and duplicate detection. Springer Science & Business Media; 2012.
20. Bozkurt, Özlem, De Boer, Anthonius, Grobbee, Diederik E., et al. Variation in renin--angiotensin system and salt-sensitivity genes and the risk of diabetes mellitus associated with the use of thiazide diuretics. American journal of hypertension. 2009; 22(5):545–551. [PubMed: 19247266]
21. Pacheco, Antonio G., Saraceni, Valeria, Tuboi, Suely H., et al. Validation of a hierarchical deterministic record-linkage algorithm using data from 2 different cohorts of human immunodeficiency virus-infected persons and mortality databases in Brazil. American journal of epidemiology. 2008; 168(11):1326–1332. [PubMed: 18849301]
22. Tromp, Miranda, Ravelli, Anita C., Bonsel, Gouke J., Hasman, Arie, Reitsma, Johannes B. Results from simulated data sets: probabilistic record linkage outperforms deterministic record linkage. Journal of clinical epidemiology. 2011; 64(5):565–572. [PubMed: 20952162]
23. Charles, Dustin, King, Jennifer, Patel, Vaishali, Furukawa, Michael F. Adoption of electronic health record systems among US non-federal acute care hospitals: 2008–2012. Office of the National Coordinator for Health Information Technology; 2013.
24. Ananthakrishnan, Ashwin N., Cai, Tianxi, Savova, Guergana, et al. Improving case definition of Crohn’s disease and ulcerative colitis in electronic medical records using natural language processing: a novel informatics approach. Inflammatory bowel diseases. 2013; 19(7):1411. [PubMed: 23567779]
25. Lin, Chen, Karlson, Elizabeth W., Canhao, Helena, et al. Automatic prediction of rheumatoid arthritis disease activity from the electronic medical records. 2013
26. McCarty, Catherine A., Chisholm, Rex L., Chute, Christopher G., et al. The eMERGE Network: a consortium of biorepositories linked to electronic medical records data for conducting genomic studies. BMC medical genomics. 2011; 4(1):1–13. [PubMed: 21208432]
27. Chen, Feng, Dow, Michelle, Ding, Sijie, et al. PREMIX: PRivacy-preserving EstiMation of Individual admiXture. American Medical Informatics Association Annual Symposium. 2016
28. Yu, Fei, Rybar, Michal, Uhler, Caroline, Fienberg, StephenE. Differentially-Private Logistic Regression for Detecting Multiple-SNP Association in GWAS Databases. In: Domingo-Ferrer, Josep, editor. Privacy in Statistical Databases. Springer International Publishing, Cham; 2010. p. 170-184.
29. Yu, Fei, Fienberg, Stephen E., Slavković, Aleksandra B., Uhler, Caroline. Scalable privacy-preserving data sharing methodology for genome-wide association studies. Journal of biomedical informatics. 2014; 50(50C):133–141. [PubMed: 24509073]
30. Uhler, Caroline, Slavkovic, Aleksandra B., Fienberg, Stephen E. Privacy-preserving data sharing for genome-wide association studies. Journal of Privacy and Confidentiality. 2013; 5(1):137–166. [PubMed: 26525346]
31. Yu, Fei, Ji, Zhanglong. Scalable Privacy-Preserving Data Sharing Methodology for Genome-Wide Association Studies: An Application to iDASH Healthcare Privacy Protection Challenge. BMC Medical Informatics and Decision Making. 2014; 14(Suppl 1):S3. [PubMed: 25521367]
32. Wang, Shuang, Mohammed, Noman, Chen, Rui. Differentially private genome data dissemination through top-down specialization. BMC medical informatics and decision making. 2014; 14(Suppl 1):S2. [PubMed: 25521306]
33. Dwork, Cynthia. Differential privacy. International Colloquium on Automata, Languages and Programming. 2006; 4052, d:1–12.
34. McSherry, Frank, Talwar, Kunal. 48th Annual IEEE Symposium on Foundations of Computer Science (FOCS’07). IEEE; 2007. Mechanism Design via Differential Privacy; p. 94-103.

35. Mohammed, Noman, Jiang, Xiaoqian, Chen, Rui, Fung, Benjamin CM., Ohno-Machado, Lucila. Privacy-preserving heterogeneous health data sharing. Journal of the American Medical Informatics Association : JAMIA. 2013; 20(3):462–469. [PubMed: 23242630]

36. Wang, Shuang, Zhang, Yuchen, Dai, Wenrui, et al. HEALER: homomorphic computation of ExAct Logistic rEgRession for secure rare disease variants analysis in GWAS. Bioinformatics. 2016; 32(2):211–218. [PubMed: 26446135]

37. Frederick, Robert. Core Concept: Homomorphic encryption. Proceedings of the National Academy of Sciences. 2015; 112(28):8515–8516.

38. Fontaine, Caroline, Galand, Fabien. A Survey of Homomorphic Encryption for Nonspecialists. EURASIP Journal on Information Security. 2007:1–15.

39. Gjøsteen, Kristian. A New Security Proof for Damgard’s ElGamal. Topics in Cryptology - CT-RSA. 2006:150–158.

40. Gentry C, Halevi S. Implementing gentry’s fully-homomorphic encryption scheme. Advances in Cryptology-EUROCRYPT. 2011:129–148.

41. Brakerski, Zvika, Gentry, Craig, Vaikuntanathan, Vinod. Proceedings of the 3rd Innovations in Theoretical Computer Science Conference on - ITCS ’12. ACM Press; 2012. (Leveled) fully homomorphic encryption without bootstrapping; p. 309-325.

42. Zhang, Yuchen, Dai, Wenrui, Jiang, Xiaoqian, Xiong, Hongkai, Wang, Shuang. FORESEE: Fully Outsourced secuRe gEnome Study basEd on homomorphic Encryption. BMC Med Inform Decis Mak. 2015; 15(Suppl 5):S5.

43. Lauter, Kristin, López-Alt, Adriana, Naehrig, Michael. Private computation on encrypted genomic data; 14th Privacy Enhancing Technologies Symposium, Workshop on Genome Privacy; 2014.

44. Zhang, Yuchen, Dai, Wenrui, Wang, Shuang, et al. SECRET: Secure Edit-distance Computation over homomoRphic Encrypted daTa; 5th Annual Translational Bioinformatics Conference;

45. Kim, Miran, Lauter, Kristin. Private Genome Analysis through Homomorphic Encryption. BMC medical informatics and decision making. 2015; 15(Suppl 5):S3. Suppl 5.

46. Chen, Feng, Cheng, Samuel, Mohammed, Noman, Wang, Shuang, Jiang, Xiaoqian. PRECISE: PRivacy-preserving cloud-assisted quality improvement service in healthcare; 2014 8th International Conference on Systems Biology (ISB); 2014. p. 176-183.

47. Constable, Scott, Tang, Yuzhe, Wang, Shuang, Jiang, Xiaoqian, Chapin, Steve. Privacy-Preserving GWAS Analysis on Federated Genomic Datasets. BMC Med Inform Decis Mak. 2015; 15(Suppl 5):S2.

48. Shi, Haoyi, Wang, Shuang, Dai, Wenrui, Tang, Yuzhe, Jiang, Xiaoqian, Ohno-Machado, Lucila. Secure Multi-pArty Computation Grid LOgistic REgression (SMAC-GLORE). BMC medical informatics and decision making. 2016; 16(Suppl 3):89. [PubMed: 27454168]

49. Chen, Feng, Mohammed, Noman, Wang, Shuang, He, Wenbo, Cheng, Samuel, Jiang, Xiaoqian. Proceedings of the 6th ACM Conference on Bioinformatics, Computational Biology and Health Informatics - BCB ’15. ACM Press; 2015. Cloud-assisted distributed private data sharing; p. 202-211.

50. Yao, Andrew C. 23rd Annual Symposium on Foundations of Computer Science (sfcs 1982). IEEE; 1982. Protocols for secure computations; p. 160-164.

51. Choi, Seung Geol, Hwang, Kyung-Wook, Katz, Jonathan, Malkin, Tal, Rubenstein, Dan. Topics in Cryptology--CT-RSA 2012. Springer; 2012. Secure multi-party computation of boolean circuits with applications to privacy in on-line marketplaces; p. 416-432.

52. Wu, Yuan, Jiang, Xiaoqian, Kim, Jihoon, Ohno-Machado, Lucila. Grid Binary LOgistic REgression (GLORE): building shared models without sharing data. Journal of the American Medical Informatics Association : JAMIA. 2012; 2012(5):758–764.

53. Wang, Shuang, Jiang, Xiaoqian, Wu, Yuan, Cui, Lijuan. EXpectation Propagation LOgistic REgRession (EXPLORER): Distributed Privacy-Preserving Online Model Learning. Journal of Biomedical Informatics. 2013; 46(3):1–50. [PubMed: 23219718]

54. Li, Yong, Jiang, Xiaoqian, Wang, Shuang, Xiong, Hongkai, Ohno-Machado, Lucila. VERTIcal Grid lOgistic regression (VERTIGO). J Am Med Inform Assoc. 2016; 23(3):570–579. [PubMed: 26554428]

55. Lu, Chia-Lun, Wang, Shuang, Ji, Zhanglong, et al. WebDISCO: a Web service for DIStributed COx model learning without patient-level data sharing; Translational Bioinformatics Conference (TBC); 2014.

56. Jiang, Wenchao, Li, Pinghao, Wang, Shuang, et al. WebGLORE: a web service for Grid LOgistic REgression. Bioinformatics (Oxford, England). 2013; 29(24):3238–3240.

57. Zhang, Yihua, Blanton, Marina, Almashaqbeh, Ghada. Secure Distributed Genome Analysis for GWAS and Sequence Comparison Computation. BMC Med Inform Decis Mak. 2015; 15(Suppl 5):S4.

58. Jiang, Xiaoqian, Zhao, Yongan, Wang, Xiaofeng, et al. A community assessment of privacy preserving techniques for human genomes. BMC medical informatics and decision making. 2014; 14(Suppl 1):S1. Suppl 1. [PubMed: 25521230]

59. iDASH Genome Privacy Protection Challenge Workshop. 2015. Retrieved March 24, 2015 from http://www.humangenomeprivacy.org/2015/

60. To Keep It Safe and Sound | GenomeWeb. Retrieved April 13, 2015 from https://www.genomeweb.com/scan/keep-it-safe-and-sound

61. Hayden, Erika Check. Cloud cover protects gene data. Nature. 2015; 519(7544):400–401. [PubMed: 25810184]

62. Sankararaman, Sriram, Obozinski, Guillaume, Jordan, Michael I., Halperin, Eran. Genomic privacy and limits of individual detection in a pool. Nature genetics. 2009; 41(9):965–967. [PubMed: 19701190]

63. Theis, Mary Kay, Reid, Robert J., Chaudhari, Monica, et al. Case study of linking dental and medical healthcare records. The American journal of managed care. 2010; 16(2):e51–e56. [PubMed: 20148610]

64. Pasquali, Sara K., Jacobs, Jeffrey P., Shook, Gregory J., et al. Linking clinical registry data with administrative data using indirect identifiers: implementation and validation in the congenital heart surgery population. American heart journal. 2010; 160(6):1099–1104. [PubMed: 21146664]

65. McCoy, Sandra I., Jones, Bill, Leone, Peter A., et al. Variability of the date of HIV diagnosis: a comparison of self-report, medical record, and HIV/AIDS surveillance data. Annals of epidemiology. 2010; 20(10):734–742. [PubMed: 20620077]

66. Weber, Susan C., Lowe, Henry, Das, Amar, Ferris, Todd. A simple heuristic for blindfolded record linkage. Journal of the American Medical Informatics Association. 2012; 19:e1, e157–e161. [PubMed: 22718034]

67. Gorelick, Marc H., Knight, Stacey, Alessandrini, Evaline A., et al. Lack of agreement in pediatric emergency department discharge diagnoses from clinical and administrative data sources. Academic Emergency Medicine. 2007; 14(7):646–652. [PubMed: 17554009]

68. Lyons, Ronan A., Jones, Kerina H., John, Gareth, et al. The SAIL databank: linking multiple health and social care datasets. BMC Medical Informatics and Decision Making. 2009; 9(1):3. [PubMed: 19149883]

69. Jurczyk P, Lu JJ, Xiong L, Cragan JD, Correa A. FRIL: A Tool for Comparative Record Linkage. AMIA 2008 Annual Symposium. 2008

70. Durham, Elizabeth, Kantarcioglu, Murat, Xue, Yuan, et al. Composite Bloom filters for secure record linkage. Knowledge and Data Engineering, IEEE Transactions on. 2014; 26(12):2956–2968.

71. Bonomi, Luca, Xiong, Li, Chen, Rui, Fung, Benjamin CM. Frequent grams based embedding for privacy preserving record linkage; ACM International Conference on Information and Knowledge Management; 2012. p. 1597-1601.

72. Mo, Huan, Thompson, William K., Rasmussen, Luke V., et al. Desiderata for computable representations of electronic health records-driven phenotype algorithms. Journal of the American Medical Informatics Association : JAMIA. 2015; 22(6):1220–1230. [PubMed: 26342218]

73. Al-Lawati, Ali, Lee, Dongwon, McDaniel, Patrick. Blocking-aware private record linkage; Proceedings of the 2nd international workshop on Information quality in information systems; 2005. p. 59-68.

74. Evangelista, Luiz Osvaldo, Cortez, Eli, da Silva, Altigran Soares, Meira, Wagner, Jr. Adaptive and Flexible Blocking for Record Linkage Tasks. JIDM. 2010; 1(2):167–182.

75. Inan, A., Kantarcioglu, M., Bertino, E., Scannapieco, M. A Hybrid Approach to Private Record Linkage; Proc. of the Int'l Conf. on Data Engineering; 2008.

76. He, Xiaoyun, Vaidya, Jaideep, Shafiq, Basit, Adam, Nabil, White, Tom. Privacy Preserving Integration of Health Care Data. International Journal of Computational Models and Algorithms in Medicine (IJCMAM). 2010; 1(2):22–36.

77. Kho, Abel N., Cashy, John P., Jackson, Kathryn L., et al. Design and implementation of a privacy preserving electronic health record linkage tool in Chicago. Journal of the American Medical Informatics Association. 2015 ocv038.

78. Adam, Nabil, White, Tom, Shafiq, Basit, Vaidya, Jaideep, He, Xiaoyun. Privacy preserving integration of health care data. AMIA Annual Symposium proceedings. 2007:1. [PubMed: 18693786]

79. Bonomi, Luca, Xiong, Li, Chen, Rui, Fung, Benjamin CM. Privacy Preserving Record Linkage via grams Projections. CoRR. 2012 **abs/1208.2**.

80. Erlich, Yaniv, Narayanan, Arvind. Routes for breaching and protecting genetic privacy. Nature reviews. Genetics. 2014; 15(6):409–421.

81. 45 C.F.R. § 46.101(b)(4).

82. Wang, Shuang, Jiang, Xiaoqian, Fox, Dov, Ohno-Machado, Lucila. Preserving genome privacy in research studies. In: Loukide, G., Gkoulalas-Divanis, A., editors. Medical Data Privacy Handbook. Springer; 2015.

83. Grady, Christine, Eckstein, Lisa, Berkman, Ben, et al. Broad consent for research with biological samples: Workshop conclusions. The American Journal of Bioethics. 2015; 15(9):34–42.

84. Beskow, Laura M., Dean, Elizabeth. Informed consent for biorepositories: assessing prospective participants' understanding and opinions. Cancer Epidemiology Biomarkers & Prevention. 2008; 17(6):1440–1451.

85. Pereira, Stacey, Gibbs, Richard A., McGuire, Amy L. Open access data sharing in genomic research. Genes. 2014; 5(3):739–747. [PubMed: 25178093]

86. Stein, Dorit T., Terry, Sharon F. Reforming biobank consent policy: a necessary move away from broad consent toward dynamic consent. Genetic Testing and Molecular Biomarkers. 2013; 17(12): 855–856. [PubMed: 24283583]

87. Edwards, Sarah JL. Research participation and the right to withdraw. Bioethics. 2005; 19(2):112–130. [PubMed: 15943021]

88. Helgesson, Gert, Johnsson, Linus. The right to withdraw consent to research on biobank samples. Medicine, Health Care and Philosophy. 2005; 8(3):315–321.

89. Eriksson, Stefan, Helgesson, Gert. Potential harms, anonymization, and the right to withdraw consent to biobank research. European Journal of Human Genetics. 2005; 13(9):1071–1076. [PubMed: 15986039]

90. Health Insurance Portability and Accountability Act (HIPAA). Retrieved from http://www.hhs.gov/ocr/hipaa

91. Slaughter, Louise. Genetic Information Nondiscrimination Act of 2008. HeinOnline; 2008.

92. Robinson, Neil, Graux, Hans, Botterman, Maarten, Valeri, Lorenzo. Review of EU data protection directive: summary. Information Commissioner's Office. 2009

93. Mostert, Menno, Bredenoord, Annelien L., Biesaart, Monique CIH., van Delden, Johannes JM. Big Data in medical research and EU data protection law: challenges to the consent or anonymise approach. European Journal of Human Genetics. 2015; 1

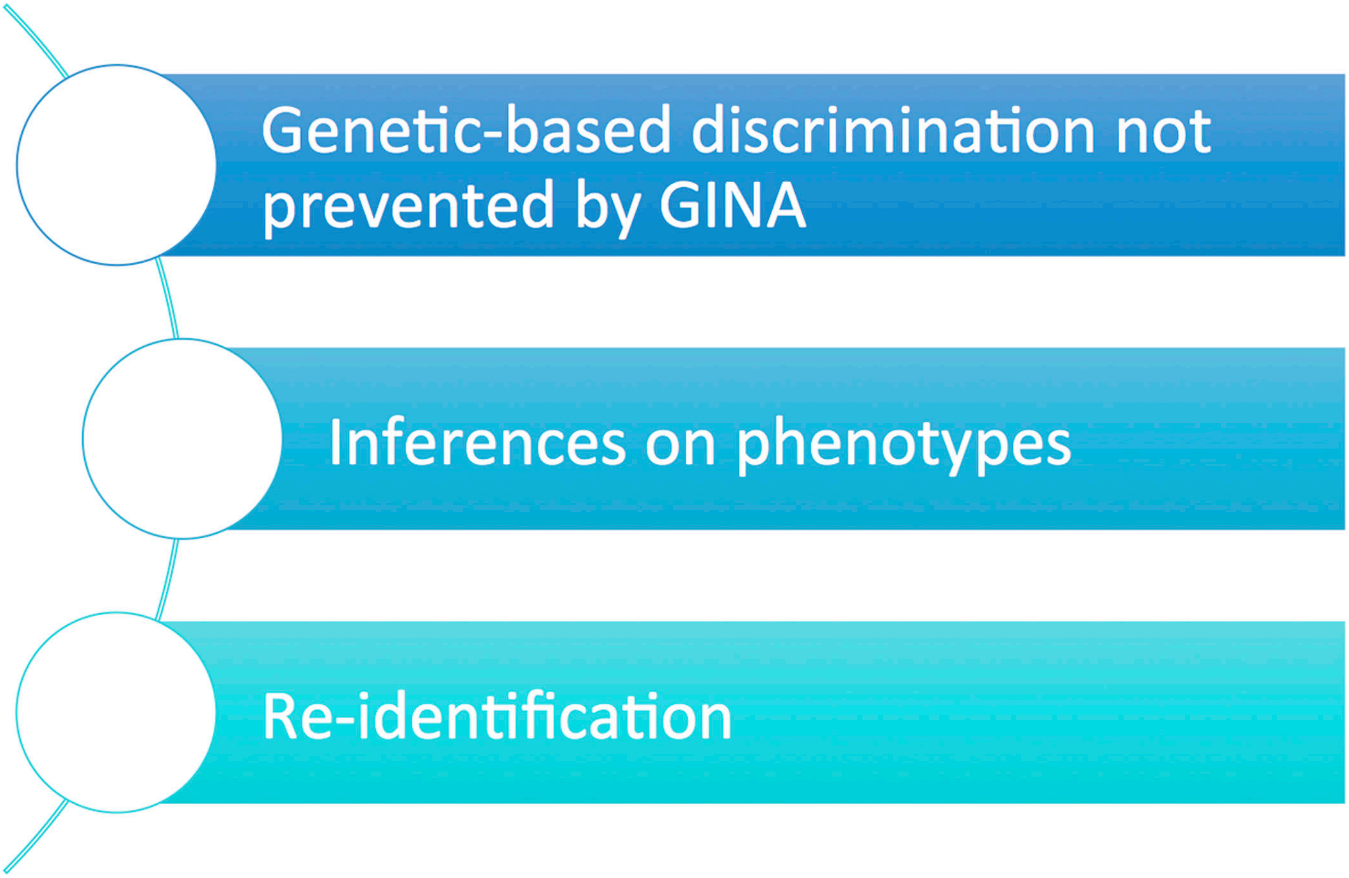
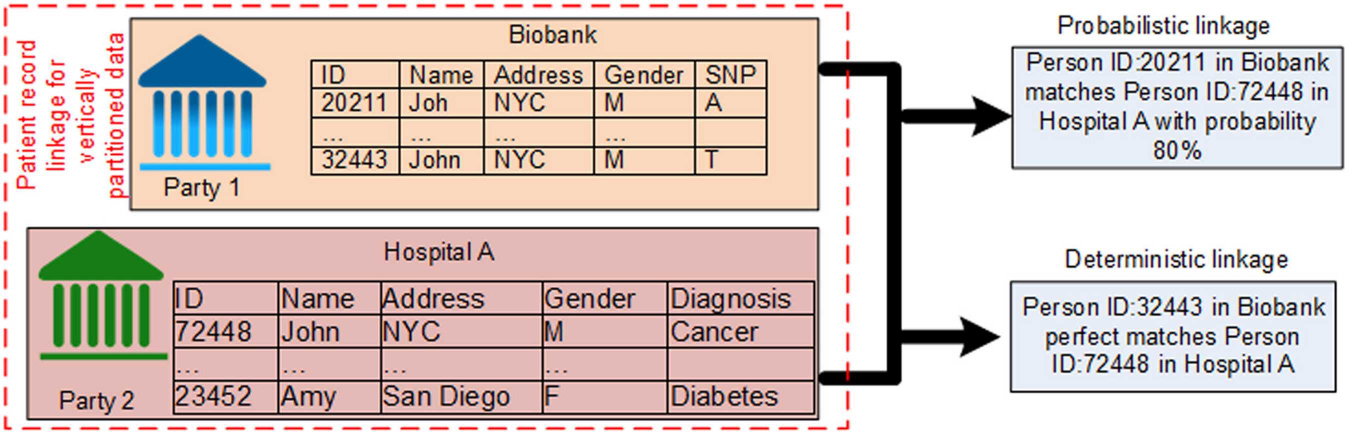


Figure 1.

Some privacy considerations in disclosing genomes and clinical data. GINA, Genetic Information Nondiscrimination Act.



ID	Name	Address	Gender	SNP
20211	Joh	NYC	M	A
…	…	…	…	
32443	John	NYC	M	T

ID	Name	Address	Gender	Diagnosis
72448	John	NYC	M	Cancer
…	…	…	…	
23452	Amy	San Diego	F	Diabetes

Person ID:20211 in Biobank matches Person ID:72448 in Hospital A with probability 80%

Person ID:32443 in Biobank perfect matches Person ID:72448 in Hospital A

Figure 2.

Examples of patient record-linkage systems for vertically partitioned data between a hospital and a biobank.