On the Security of Biobanks in Biomedical Informatics Research

Tos Chan
Tufts University, Fall 2019

Abstract

Biomedical informatics research has become increasingly reliant on biobanks. Biobanks are data repositories for which the primary form of data contained are biological datasets, such as genomes, transcriptomes, and proteomes. While the analysis of this data is incredibly useful for the advancement of medical technology, there are many security concerns with allowing researchers access to large amounts of personal health data. Upholding the security of biobanks is crucial in practicing ethical research, maintaining trust with bank donors, and preventing patients from the dangers of having their privacy violated. This paper will first begin by exploring the security regulations and patient privacy compliance for biobanks. It will then cover common vulnerabilities and security risks in data banks. The paper will also discuss notable security techniques and their effectiveness. Finally, the paper will propose improvements in policy and software implementation for increased protection of patient privacy and data security. By working to secure biobanks, participant trust is increased, donation to banks is encouraged, and scientists are allowed continued access to a plethora of data for high-impact research.

I. Introduction

With the increasing ability to store and analyze large amounts of data, biomedical informatics research has become largely dependent on biobanks. These databases are crucial in the development of new drugs and in predicting and identifying root causes of illness. Nevertheless, the information stored in these repositories is highly sensitive, oftentimes containing DNA/RNA samples. DNA is the most personal form of data as it is the unique identifier of an individual. Genetic data is capable of identifying family relations, elucidating regional ancestry, predicting appearance, health, and disease, and assisting forensic criminal investigations [1]. Furthermore, exposure of this data can lead to discrimination from health insurance providers and employers. There are some limited protections offered in the United States, but more exhaustive policy is needed [2].

These concerns present a crucial question: how can biobank data accessibility be maintained without compromising the anonymity of donors [3], [4]? This question is especially pertinent considering the newfound promotion to create larger biobanks for more comprehensive research. It is imperative to ensure the security of these biobanks through both software techniques and legislation to sufficiently protect the privacy of personal health data and conduct safe and ethical research.

II. To the Community

To sustain high-impact biomedical research, we must keep information stored in biobanks accessible to researchers. These data banks can only function through the trust of its donors and adherence to privacy protection policies. Thus, it is of utmost importance that the security and privacy of participants and their data are prioritized when creating and maintaining these biobanks. Furthermore, awareness of the security issues that plague these data banks can help can inform potential donors and help maintainers and policymakers better protect your health data.

III. Background

A. *Definition of Biobanks*

Biobanks are repositories that collect and store biological material ranging from organic matter like tissue samples to biometrics like DNA/RNA samples. Additionally, biobanks contain annotations associated with the sample data. For example, they might note age, sex, ethnicity, and medical history alongside a participant's sample [4]. The inclusion of these annotations allows investigators to perform more statistical analyses, but it can also increase risks if donor data is identified since more personal information can be linked back to them.

B. *Biobanks in Research*

Biobanks are now heavily relied upon for biomedical informatics research. For example, bioinformaticians can perform genome-wide association studies (GWAS) to identify particular genetic markers that are linked to inherited conditions [5]. The heuristics needed to analyze the data are not new but require large amounts of data due to the weakness of associations. The ability to conduct these tests has been made possible through increased access to the sequenced genomes from thousands of individuals. Similarly, computational biologists are using neural nets to predict whether genetic variants are benign or pathogenic. Again, this research was only possible due to the existence of biobanks [5]. Biobanks are also frequently used for precision medicine and drug development purposes. Instead of developing drugs that curb symptoms, scientists can now develop drugs that work in accordance with a patient's genes [5]. Evidently, biobanks are vital to the advancement of biomedical informatics research.

C. *Biobank Security Concerns*

Given the role of biobanks in research, there are unique privacy concerns and risks that should be considered. Specifically, the privacy of samples is precarious in research contexts because biobanks are not usually conducting research but helping to facilitate it. The privacy of donors is covered by general privacy legislation as well as research standards. However, this burden of compliance is placed on research institutions working with biobanks but not on the data banks themselves. Thus, promises of privacy that biobanks may offer to participants are questionable. [2] This is even more concerning when accounting for informed consent. If data banks are not held sufficiently accountable for data protection, are participants being falsely assured [4]?

IV. Regulations

A. *Privacy Laws*

Numerous policies are designed to protect personal health information and sample data donated for research purposes. First, there are general privacy laws, such as the Fourth Amendment and the Privacy Act of 1974. The Fourth Amendment grants the "right to privacy" but only applies to donors with samples hosted by government-run biobanks or researchers employed by the United States government [4, p. 114]. These protections are limited and often prioritize the government's interest to disclose the information over an individual's right to privacy. The Privacy Act of 1974 is one of the oldest federal privacy laws. Similarly, its protections only apply to biobanks maintained by federal agencies. It limits the dissemination of records without the consent of the individual, though there are twelve statutory exceptions. Two of these exceptions can apply to biobanks: records collected for research purposes could fall under "routine uses" or "statistical research" [4, p. 115].

Additional privacy protections include the Health Insurance Portability and Accountability Act of 1996 (HIPAA) and the Genetic Information Nondiscrimination Act of 2008 (GINA). HIPAA was established as an insurance statute and was not intended to rigorously protect privacy. When Congress failed to push further privacy policy, they charged the Department of Health and Human Services (HSS) with the responsibility. HHS added the Privacy Rule under GINA, which treats genetic information

"under the ordinary protections of the HIPAA Privacy Rule" [1, p. 12]. This rule restricts covered entities from using or disclosing personal health information (PHI), which includes genetic data, except when data is anonymized (by removing specific identifiers) or when the subject gives their written consent to have their data disclosed. There are also HIPAA Public Purpose exceptions as well, which allow institutions to disclose PHI without consent [1]. Modifications to HIPAA in 2013 also eased the burden of privacy protection on research by streamlining the participant consent process [6].

B. Research Laws

There also exist research-specific privacy laws, such as the Federal Policy for the Protection of Human Subjects (Common Rule) and the Food and Drug Administration (FDA)'s human subject protection regulations (Title 21, Code of Federal Regulations, Parts 50 and 56). The Common Rule grants protections to human subjects for government-funded research in the United States. This also applies to other institutions that receive Federalwide Assurance (FWA) [4]. Thus, biobanks may be regulated under the Common Rule if the research is conducted by a government-affiliated entity. Even in these cases, however, the rule may not apply to data banks because they simply collect and archive data -- which is not deemed as research. Only activity with an investigation aspect falls under the Common Rule [4]. The FDA's regulations are quite similar to the Common Rule but place the responsibility of enforcement on local Institutional Review Boards. Both the Common Rule and FDA regulations require the anonymization of genetic information for purposes of disclosure [1].

The above policy only applies to the United States and is further complicated by differences between state law protections [1]. Other countries offer different standards of protection. Countries in the European Union (EU) adopted the European Union General Data Protection Regulation (GDPR) in 2016 [6]. These differences in policies have created issues when sharing data from biobanks across borders [7]. The EU/US Safe Harbor Provisions were created, in turn, to guarantee safe data transfer from the EU to the United States for companies but does not apply to nonprofits, which includes academic researchers. Furthermore, it does not apply to the direct transfer of data between colleagues overseas, but it does apply to the posting of data to US-owned cloud computer sources [4].

C. Issues in Policy

Although there are multiple privacy policies in place, the existing legislation was not written with the intention to protect participants of biobanks [4]. This is reflected in the inconsistent manners in which biobank-related privacy is upheld. For biobanks, determining when protections are offered is specific to a unique set of functions and relationships [8]. For example, the Common Rule applies to government-funded research while the Privacy Rule applies to covered entities, essentially institutions associated with health care claims. Furthermore, these protections offer different standards. Although the Common Rule, the Privacy Rule, and FDA regulations enforce the de-identification of genetic information, the standards of what is considered identifiable vary. The standards are the strictest in the FDA and least strict in the Common Rule [4]. The inconsistencies of coverage create a level of complexity that makes biobank-based research "legally complicated, unnecessarily expensive, and frequently delayed by issues of privacy" [4, p. 121]. Despite multiple privacy policies, there remains a lack of comprehensive legislation that applies directly to biobanks as well as inconsistent enforcement of privacy protection. This calls into question whether existing legislation effectively protects participants from security risks.

V. Technical Security Risks

A. Identity Tracing Attacks

Identity tracing attacks aim to uniquely identify the owner of an anonymized DNA sample with the use of quasi-identifiers embedded in the dataset. These identifiers, such as age, sex, and zip code, can

be acquired through social media or public record search portals, eg. PeopleSmart and FindOutTheTruth [9]. This has been seen in the Personal Genome Project, in which 30% of participants were later dentified using birth dates, sex, and zip codes. [10] Another method used to conduct these attacks is through triangulation with Y chromosomes. Studies have demonstrated that surnames can be inferred by combining metadata with Y chromosomes obtained through websites like Ysearch or SMGF [9].

In response to these issues, the Global Alliance for Genomics and Health (GA4GH) created the beacon, a project designed to share genomic data securely. Researchers may pose questions, such as "Do you have an allele at a specific position in a genome," and the beacon can answer "yes" or "no" [9, p. 4]. This web-based service was created to prevent identity tracing attacks, but recent studies have shown that they are still vulnerable to re-identification attempts. If attackers have background knowledge of specific allele frequencies in the beacon, it gives adversaries higher re-identification power by correctly inferring alleles at certain positions through multiple queries [9].

These techniques are often used in forensics. Suspects can be identified through their relative's DNA. Long-range familial searches can be conducted with third-party services like GEDmatch [9]. This was performed in the identification of the Golden State Killer [11]. In fact, the FBI has a database known as CODIS (Combined DNA Index System), which is designed to profile suspects using DNA profiling.

B. Attribute Disclosure Attacks

Attribute disclosure attacks are those in which an adversary predicts sensitive attributes of victims, such as "phenotypes, disease association, and drug abuse" [9, p. 6]. These cases are distinct from identity tracing attacks as the identity of the DNA is known. Instead, they wish to gain information about victims that could harm them. A simple way to conduct these attacks is to take the unknown DNA and search disease-driven biobanks for a match. This would identify the donor and link them to that disease [2]. These attacks can be refined by looking at single nucleotide polymorphisms (SNPs) in biobank-enabled GWASs. In other words, attackers can check for the presence of known markers associated with a disease to attribute it to their victim [10].

Another method to perform an attribute disclosure attack is through a query-based tracker. For example, a researcher may know the identity of a male victim and want to determine if he has HIV. They can create a query that satisfies the restrictions on his identity and HIV diagnosis. If the number of results is 1, then the victim is known to have HIV. This can be prevented with lower limit thresholds, but this is not always successful. A researcher can bypass limits by performing a sequence of queries as follows: query the number of people who are female and HIV-positive, then query the number of people diagnosed with HIV who are either female or who satisfy the restrictions based on the male victim's identity. They can subtract the number from that of the previous query to determine if the victim has HIV [12].

C. General Database Security

General database security risks also apply to biobanks. For example, the database can still be subject to SQL injection. Even in NoSQL systems, injection attacks are possible. These injections also have greater potential impact due to using a procedural rather than declarative language like SQL [13]. These NoSQL methods are more popularly used by researchers and subject to many security issues [14]. In many NoSQL systems, authorization and authentication are not enabled by default, and there is insufficient encryption for data files and attributes. Client-server authentication is also weak—oftentimes using HTTP basic, which is vulnerable to man-in-the-middle attacks [15]. Although the data is generally accessible to many, it is through proper authorization of researchers. When biobanks do not practice basic database security methods, it allows for more individuals to perform the aforementioned attacks.

VI. Techniques for Protection

A. *Cryptographic Techniques*

The International Society for Biological and Environmental Repositories (ISBER) recommends using cryptographic primitives [16]. Cryptographic primitives are low-level algorithms used to create cryptographic systems that provide information security. An example is homomorphic encryption (HE), which allows computation on its ciphertext [9]. Homomorphic encryption is helpful when publicizing data to a cloud or sharing data, keeping the data private but allowing analyses to be conducted. In the case of biobank-based research, maintainers could encrypt SNPs and allow researchers to conduct GWASs. The computational cost from HE with GWAS is immense, with a multi-party approach requiring 30 years of computation time. However, recent work proposed a new technique to reduce the overhead to just a few days for computation and data transfer [9].

Secure cryptographic hardware (SCH) should also be considered when working to share and compute genomic information. SCH employs encryption through processor instruction set and offers unique benefits. SCH is both more efficient and versatile than HE as it is not limited to only multiplication and addition but can compute arbitrary functions. Additionally, it is more resistant to tampering than software. It should be noted, however, that they are restricted in memory size especially for computation of genomic data which consists of millions of nucleotides [9].

B. *Tracker Prevention Techniques*

Multiple techniques can be implemented to prevent query-based tracking. The first is simply to use tracker detection, which analyzes a series of queries to determine if inference attacks can be performed. Most detections methods are quite inefficient though. Another method is the anonymization of the dataset; this is distinct from encryption. Datasets can be anonymized by removing data that can reveal the owner of a sample (e.g. unique identifiers like social security numbers), using relative data (e.g. age instead of birth date), or using more coarse-grained data (e.g. age groups in five-year intervals). The last technique is adding statistical noise to results. Adding noise can prevent attackers from deducing individual data but still allow researchers to find relevant cases. This technique is complex since it must blur the data enough to prevent the attacker from reducing noise through multiple queries [12].

C. *Database Security Techniques*

General database security techniques are also helpful in securing personal health data. Upholding standards of authentication, authorization, and auditing are simple yet still overlooked in biobanks. Authentication can be improved by enforcing strong passwords through a minimum character length, requiring the use of special characters, and banning dictionary words [8]. In terms of authorization, using the principle of least privilege will limit the number of people accessing sensitive information. For example, the National Institute of Health uses a two-tiered role-based access control system for GWAS datasets: a restricted access area that stores the encrypted data and a public access area that contains summary statistics and allele frequency for cases and controls [10]. Additionally, biobanks can use auditing as required in HIPAA, logging the actions of researchers for each resource and using error detection to train for and address violations [8].

Furthermore, networks can be configured for secure transfer of biobank data. The Tohoku Medical Megabank (TMM) data sharing policy advises the use of closed sub-networks to share different information. They outline three separate sub-networks: one for identifiable data, one for de-identified data, and one for shared data. Each of these networks is independent and each controlled by its own firewall. Data transfer between these subnetworks should occur offline and access to the networks should be accessible through IP-VPN [17]. IP-VPN is more secure than traditional VPNs, which operate on the third (network) layer of the Open Systems Interconnection (OSI) model and are susceptible to distributed denial of service (DDoS) attacks. IP-VPNs use the second (data link) layer, traveling on remote connections, keeping data secure [18].

VII. Action Items

With the advancements in DNA sequencing technologies, biobank sample collections are expanding and more personal health data is at risk. This past year, the National Institute of Health initiated the All of Us program, which aims to create a biobank of one million participants of various ethnic backgrounds. The program was found to have vulnerabilities that would allow attackers to compromise participants' personal health information. [8] The failure to protect donor data is indicative of inadequacies in biobank security. This is unsurprising provided the lack of policy intentionally designed to regulate the security of biobanks. Current policy either applies to biobanks as an afterthought or applies solely to biobank-enabled research and not biobanks themselves. To ensure that donor data is sufficiently protected and research continues to advance, it is imperative to supplement the existing policies with more comprehensive legislation.

To address the issues of sharing data across state and country borders, an international standard should be created. The standard should eliminate the complexity created by differences in US policy. By analyzing HIPAA, FDA regulations, and the Common Rule, the new policy can encompass the strengths and effectiveness of each and unify differing definitions (such as that of de-identifiable information). Furthermore, this standard should take into account the privacy protections offered by other countries given that it will be applied internationally. There are surely protections in the GDPR that can be adopted by the standard.

Furthermore, there is little government policy outlining secure software implementations. This is not surprising provided many legislators do not have stakes in the cybersecurity community. There exist multiple guidelines put out by independent organizations, such as ISBER. However, the existence of multiple standards poses the same issues of complexity as the various privacy policies in place. Creating legal standards and enforcement mechanisms with experts from the field should be considered. Two possible options are either forming a new governing board dedicated to the enforcement of these standards or expanding local IRB responsibilities to include the review of security techniques for biobanks in their jurisdiction. These policy additions should be included in the international standard.

The creation of an international standard will hopefully elucidate any prior legal complexity and increase the efficiency and cost-effectiveness of research. These complexities have been alleviated at the expense of privacy as seen in the past. In 2015, the notice of Proposed Rulemaking (NPRM) gave stricter proposals to Common Rule but was struck down by pressure from researchers. We must not follow this precedent; concerns of bureaucracy should always be placed second to participant privacy. Prioritizing an individual's right to privacy in policy is not only becoming increasingly necessary in a world with rapidly advancing technology, but it is the only way for research to continue. Specifically, research must include informed consent and the trust of donors to sustain the growth of biobanks and research enabled by them.

VIII. Conclusion

The existing policy framework is complex and costly, ultimately slowing the process of research and failing to protect individual rights to privacy. Despite the existing methods to prevent attacks, there have still been breaches in biobank security. These institutions are not employing the proper techniques to protect biobank data. To address these issues, intentional and comprehensive policy must be amended, supplemented, and enforced. Improving the overall security of biobanks is beneficial for all stakeholders, since it encourages more donations for important research without sacrificing donors' sensitive health data. Furthermore, streamlining policy will reduce time approval costs of research, making research both more efficient and ethical.

IX. References

[1] E. W. Clayton, B. J. Evans, J. Hazel, and M. A. Rothstein, "The Law of Genetic Privacy: Applications, Implications, and Limitations," Journal of Law and the Biosciences, vol. 6, no. 1, pp. 1–36, May 2019.

[2] S. M. Suter, "GINA at 10 years: the battle over 'genetic information' continues in court," *Journal of Law and the Biosciences*, vol. 5, no. 3, pp. 495–526, Jan. 2018.

[3] H. Langhof, H. Kahrass, S. Sievers, and D. Strech, "Access policies in biobank research: what criteria do they include and how publicly available are they? A cross-sectional study," *European Journal of Human Genetics*, vol. 25, no. 3, pp. 293–300, Dec. 2016.

[4] H. L. Harrell and M. A. Rothstein, "Biobanking Research and Privacy Laws in the United States," *The Journal of Law, Medicine & Ethics*, vol. 44, no. 1, pp. 106–127, Mar. 2016.

[5] J. A. Diao, I. S. Kohane, and A. K. Manrai, "Biomedical informatics and machine learning for clinical genomics," *Human Molecular Genetics*, vol. 27, no. 1, pp. 29–34, May 2018.

[6] M. J. Bledsoe, "Ethical Legal and Social Issues of Biobanking: Past, Present, and Future," *Biopreservation and Biobanking*, vol. 15, no. 2, pp. 142–147, 2017.

[7] K. M. Berger and P. A. Schneck, "National and Transnational Security Implications of Asymmetric Access to and Use of Biological Data," *Frontiers in Bioengineering and Biotechnology*, vol. 7, 2019.

[8] R. Heatherly, "Privacy and Security within Biobanking: The Role of Information Technology," *The Journal of Law, Medicine & Ethics*, vol. 44, no. 1, pp. 156–160, Mar. 2016.

[9] A. M. Yakubu and Y.-P. P. Chen, "Ensuring privacy and security of genomic data and functionalities," *Briefings in Bioinformatics*, Feb. 2019.

[10] Y. Erlich and A. Narayanan, "Routes for breaching and protecting genetic privacy," *Nature Reviews Genetics*, vol. 15, no. 6, pp. 409–421, Aug. 2014.

[11] R. Winton, J. Serna, P. St. John, and B. Oreskes, "Police used consumer genealogical websites to identify Golden State Killer suspect," *Los Angeles Times*, 26-Apr-2018.

[12] J. Eder, H. Gottweis, and K. Zatloukal, "IT Solutions for Privacy Protection in Biobanking," Public Health Genomics, vol. 15, no. 5, pp. 254–262, 2012.

[13] "Testing for NoSQL injection," in *OWASP Testing Guide*, The Open Web Application Security Project , 2016.

[14] W. L. Schulz, B. G. Nelson, D. K. Felker, T. J. Durant, and R. Torres, "Evaluation of relational and NoSQL database architectures to manage genomic annotations," *Journal of Biomedical Informatics*, vol. 64, pp. 288–295, 2016.

[15] N. Gupta and R. Agrawal, “NoSQL Security,” *Advances in Computers A Deep Dive into NoSQL Databases: The Use Cases and Applications*, pp. 101–132, 2018.

[16] A. Parry-Jones, “ISBER best practices: recommendations for repositories, 4th edition,” Cryobiology, vol. 85, p. 152, 2018.

[17] T. Takai-Igarashi, K. Kinoshita, M. Nagasaki, S. Ogishima, N. Nakamura, S. Nagase, S. Nagaie, T. Saito, F. Nagami, N. Minegishi, Y. Suzuki, K. Suzuki, H. Hashizume, S. Kuriyama, A. Hozawa, N. Yaegashi, S. Kure, G. Tamiya, Y. Kawaguchi, H. Tanaka, and M. Yamamoto, “Security controls in an integrated Biobank to protect privacy in data sharing: rationale and study design,” *BMC Medical Informatics and Decision Making*, vol. 17, no. 1, Jul. 2017.

[18] “What is the Difference Between VPN and IP VPN?,” iTel, 15-Oct-2019. [Online]. Available: https://itel.com/what-is-difference-vpn-and-ip-vpn/. [Accessed: 07-Dec-2019].