Navigating ethical quandaries with the privacy dilemma of biomedical datasets

Gamze Gürsoy*
Yale University, New Haven, CT, 06511, USA
Email: gamze.gursoy@yale.edu

Megan Doerr
Sage Bionetworks, Seattle, WA, 98109, USA
Email: megan.doerr@sagebionetworks.org

John Wilbanks
Sage Bionetworks, Seattle, WA, 98109, USA
Email: john.wilbanks@sagebionetworks.org

Jennifer K. Wagner
Geisinger, Danville, PA, 17822, USA
Email: jwagner1@geisinger.edu

Haixu Tang
Indiana University Bloomington, Bloomington, IN, 47405, USA
Email: hatang@indiana.edu

Steven E. Brenner†
University of California Berkeley, Berkeley, CA, 47420, USA
Email: brenner@compbio.berkeley.edu

With decreasing cost of biomedical technologies, the scale of the genetic and healthcare data have exponentially increased and become available to wider audiences. Hence, privacy of patients and study participants has garnered the attention of researchers and regulators alike. Availability of genetic and health care information for uses not anticipated at the time of collection gives rise to privacy concerns such that people suffer dignitary harm when their data is used in ways they did not desire or intend, even if no concrete economic damage results. In this workshop, we explore the issues surrounding data use to advance human health from a privacy perspective. Broadly this field can be considered in two encompassing areas: (1) Ethics and regulation of privacy: The ethical and regulatory frames through which we can consider privacy, the existing regulations regarding privacy and what is on the horizon, and implementation of such ethical considerations for data with the new Common Rule. (2) Approaches to ensuring privacy using technology: The technologies that allow responsible use and sharing of data such as encryption and the quantification of privacy leakages in publicly available data through privacy attacks for better risk-assessment tools.

Keywords: ethics, privacy, data sharing, biomedical data, human subjects

* This work is partially supported by NIH grant U01EB023686.

† This work is partially supported by NIH grant U01EB023686, U41HG007346 and Tata Consultancy Service

Data privacy is an important topic arising from fields such as technology and medicine, and requires insights integrating many different fields such as ethics, sociology, law, political science, and forensic science. Genetic and other health data has emerged as a major focus of privacy advocates and researchers. This can be attributed to the advancement of biotechnology, the steep declines in the cost of data acquisition, and efforts to analyze such data at large scale to understand biology and enhance medical case. As a result, there is a surge of datasets that have been collected, processed, and harmonized from different sources, such as genomic data, electronic health records (EHR), and data from mHealth devices. It has been shown that in addition to the genomic data [1], high throughput molecular phenotype datasets such as functional genomic and metabolomics measurements, and microbiome measurements can be used by adversaries for re-identification purposes [2,3,4]. In addition, the emergence of EHRs with the rise of personalized medicine makes patients vulnerable to privacy breaches. These observations indicate that privacy concerns over sharing personal biological data will increase quickly with the sharing of consumer genetic data. The data collection and sharing procedures that these companies use and how these procedures are regulated call for a public discussion of privacy considerations around these new concepts.

The privacy of participants' information is a core tenant of human subject research, codified by the Declaration of Helsinki (1964) which establishes the duty of physicians involved in medical research to protect "privacy . . . and confidentiality of personal information of research subjects" [5]. The International Ethical Guidelines for Health-Related Research Involving Humans (International Ethical Guidelines) (2016) further addresses the requirements for consent for digitally-derived data and the residual privacy risks despite safeguards[6]. Of particular note, the International Ethical Guidelines specifically call out the responsibility of researchers to ask for permission (through a minimum "opt out") to use digital data for research and prohibits its use for research if the data subject objects. From a regulatory perspective, the requirements are far less plain. A multitude of intersecting agencies and regulations, with large and unexpected gaps characterizes the current state of affairs in the US. In the absence of clarity over federal agencies' jurisdictional boundaries, federal regulators have struggled to address single source data use/misuse, a problem that will be magnified several fold as datasets with pejoratively different regulatory frames are integrated for use in precision health and beyond[7,8]. In fact, many have noted the importance of transparency and accountability in data science, including including, Price, Spector-Bagdady, and colleagues, who describe "shadow health records" (i.e., "collections of health data outside the health system that provide detailed pictures of individual health"), highlighting current evasive or workaround practices to data privacy restrictions[9]. These themes will be explored in depth by our panelists.

The benefit and importance of open data sharing is widely acknowledged. However, privacy concerns have led to procedures such as controlled access (e.g., dbGaP) that inhibit the access to the data by average researchers by creating bureaucratic bottlenecks and impeding integration and collaborative development. Hence the technical advances that prevent the privacy leakage while promoting data sharing are essential. This highlights the importance of cryptographic techniques that can compute on encrypted data or novel data dissemination systems that allow sharing while

protecting privacy. Another way to protect privacy of study participants and patients is studying the quantification of the prospective privacy loss before the release of the data, and permitting participants for more encompassing data sharing.

Moving forward, it will be important to find a way to address mounting privacy protection concerns in an ethical framework to ensure that individuals are protected even as their aggregated data are shared broadly enough to promote biomedical advances for everyone's health.

References

1. Gymrek M, McGuire AL, Golan D, Halperin E, Erlich Y. Identifying personal genomes by surname inference. Science, 2013;339(6117):321-324
2. Harmanci A, Gerstein M. Analysis of Sensitive Information Leakage in Functional Genomics Signal Profiles through Genomic Deletions. Nature Communications, 2018; 9 (1), 2453
3. Gürsoy G, Harmanci A, Green M, Navarro F, Gerstein M. Sensitive information leakage from functional genomics data: Theoretical quantifications & practical file formats for privacy preservation, 2018, Biorxiv
4. Franzosa EA, Huang K, Meadow JF, Gevers D, Lemon KP, Bohannan BJ, Huttenhower C.Identifying personal microbiomes using metagenomic codes. Proc Natl Acad Sci U S A.112(22):E2930-8 (2015)
5. World Medical Association, Declaration of Helsinki, General Principles, (1964).
6. Council for International Organizations of Medical Sciences, International Ethical Guidelines for Health-Related Research Involving Humans, Guideline 22 (4th ed. 2016).
7. Fair, L. "FTC's $5 billion Facebook settlement: Record-breaking and history-making." Federal Trade Commission. Last updated June 24, 2019. https://www.ftc.gov/news-events/blogs/business-blog/2019/07/ftcs-5-billion-facebook-settlement-record-breaking-history (Last accessed October 7, 2019).
8. Coldewey, D. "9 reasons the Facebook FTC settlement is a joke." TechCrunch Last updated June 24, 2019. https://techcrunch.com/2019/07/24/9-reasons-the-facebook-ftc-settlement-is-a-joke/ (Last accessed October 7, 2019).
9. Price, WN II, Kaminski ME, Minsenn T, Spector-Bagdady K. "Shadow health records meet new data privacy laws." Science. 2019; 363(6426):448-450. PMCID: PMC6417878.