Online Ethics Center
FOR ENGINEERING AND SCIENCE

Case: Big Data & Genetic Privacy: Re-identification of Anonymized Data

Author(s)

Valerie Racine

Year

2017

Description

In 2013, Melissa Gymrek, Amy McGuire, David Golan, Eran Halperin, and Yaniv Erlich published an article describing how they re-identified almost 50 individuals from "anonymized" data in a genomic database from the 1000 Genomes Project. Their goal was to show the vulnerability of genomic databases to this sort of manipulation. The case opens discussion of sharing genomic data and protecting privacy.

Body

In a study published in *Science* in 2013, researchers outlined how they were able to re-identify almost 50 individuals from “anonymized” data in a genomic database from the 1000 Genomes Project (Gymrek *et al.* 2013). Their intention was to “demonstrate end-to-end identification of individuals with only public information,” using simple computational search tools and an internet connection (Gymrek *et al.* 2013, 321).

In general, researchers can re-identify specific individuals or small groups by using “quasi-identifiers” to cross-reference certain data included in the genetic databases that are also available in other databases (Kupersmith 2013). These “quasi-

identifiers" can come from a variety of public and non-public databases, such as hospital data, ICD-9 codes,[1] social security database, vehicular databases, voter registration lists, house sales, and other public records' search engines (Kupersmith 2013). To re-identity anonymized data, then, researchers can use computational approaches to match data from a candidate anonymized database with the data from one or more reference databases, using their shared elements such as zip codes.

Gymrek *et al*. used data from individuals who had been sequenced for the Center for Study of Human Polymorphisms (CEPH) family collection, and were stored in the 1000 Genome Project. The participants of the research were informed that the database provided broad and open access to the data for genomic analyses, and that there was a slight risk that re-identification was possible. Privacy was not promised to the participants. Still, it was assumed that the risk of re-identification was low (Rodriguez *et al.* 2013, 275).

In their study, the researchers leveraged information about patrilineal relations from databases to re-identify individuals by surnames. They used sequence data to identify single nucleotide polymorphisms (SNPs) on the Y chromosome in the genomes of individuals (Gymrek *et al* 2013, 323; Kupersmith 2013). These SNPs, referred to as Y-STR (short tandem repeats) markers, are used to identify patrilineal lineages. They then used this information to search databases which included the surnames of 40,000 individuals and their pedigrees. Next, they matched that information with other public sources of information from the National Institute of General Medical Sciences (NIGMS) Human Genetic Cell Repository at the Coriell Institute. That database included information about obituaries, as well as information from the biological materials gathered for the CEPH. As a result of this search procedure, which took only a few hours to complete, the researchers were able to identify almost 50 individuals, although they did not disclose any individual names in the publication of their research.

Before the publication of the study, the researchers contacted the National Institutes of Health (NIH), whose staff members then consulted with the editors of *Science* and the staff working for the CEPH study, to discuss what to do about the privacy breach they demonstrated in their study (Rodriguez *et al.* 2013, 275-276). Changes were made to the publicly-accessible repository, including the removal of any information indicating the age of the participants. But, none of the methods that the researchers used violated the existing laws or regulations designed to protect individuals'

genetic privacy and prevent genetic discrimination.

Genomic and genetic data about individuals or groups are particularly sensitive because they can have stigmatizing consequences, such as “employment discrimination, denial of life insurance, and inappropriate marketing” (Kupersmith 2013). Consequently, this study triggered many questions about how best to ensure the privacy of research participants and promises of confidentiality and, more importantly, how to balance the competing goals of scientific research in genomics with respect for individual autonomy.

Discussion Questions

1. Should there be additional regulations restricting public access to genomic databases? If so, who may have access to them and how? Who should decide the qualifications required for researchers to gain access to databases?
2. What are the researchers’ moral responsibilities to research participants who consent to the collection and storage of their genomic sequence?
3. What are the research participants’ (and citizens’, more generally) moral responsibilities to participate in the collection and storage of genetic and genomic information in databases and consent to the sharing of that data for further genomic analyses?

Bibliography

Altman, Russ B., Ellen Wright Clayton, Isaac S. Kohane, Bradley A. Malin, and Dan M. Roden. “Data re-identification: societal safeguards.” *Science* 339, no. 6123 (2013): 1032-1033.

Angrist, Misha. “Eyes wide open: the personal genome project, citizen science and veracity in informed consent.” *Personalized medicine* 6, no. 6 (2009): 691-699.

Barocas, Solon, and Helen Nissenbaum. “Big data's end run around procedural privacy protections.” *Communications of the ACM* 57, no. 11 (2014): 31-33.

Chalmers, Don, Michael Burgess, Kelly Edwards, Jane Kaye, Eric M. Meslin, and Dianne Nicol. “Marking shifts in human research ethics in the development of biobanking.” *Public Health Ethics* (2014): phu023.

Clayton, Ellen Wright. “Ethical, legal, and social implications of genomic medicine.” *New England Journal of Medicine* 349, no. 6 (2003): 562-569.

Erlich, Yaniv, and Arvind Narayanan. “Routes for breaching and protecting genetic privacy.” *Nature Reviews Genetics* 15, no. 6 (2014): 409-421.

Gymrek, Melissa, Amy L. McGuire, David Golan, Eran Halperin, and Yaniv Erlich. “Identifying personal genomes by surname inference.” *Science* 339, no. 6117 (2013): 321-324.

Hudson, Kathy L. “Genomics, health care, and society.” *New England Journal of Medicine* 365, no. 11 (2011): 1033-1041.

Ioannidis, John PA. “Informed consent, big data, and the oxymoron of research that is not research.” *The American Journal of Bioethics* 13, no. 4 (2013): 40-42.

Kaye, Jane, Paula Boddington, Jantina de Vries, Naomi Hawkins, and Karen Melham. “Ethical implications of the use of whole genome methods in medical research.” *European Journal of Human Genetics* 18, no. 4 (2010): 398-403.

Kupersmith, Joel. “The privacy conundrum and genomic research: re-identification and other concerns.” *Health Affairs Blog.* September 11, 2013. Accessed October 19, 2016. http://healthaffairs.org/blog/2013/09/11/the-privacy-conundrum-and-genomic-research-re-identification-and-other-concerns/

Larson, Eric B. “Building trust in the power of ‘big data’ research to serve the public good.” *JAMA* 309, no. 23 (2013): 2443-2444.

Lunshof, Jeantine E., Ruth Chadwick, Daniel B. Vorhaus, and George M. Church. “From genetic privacy to open consent.” *Nature Reviews Genetics* 9, no. 5 (2008): 406-411.

Master, Zubin, Lisa Campo-Engelstein, and Timothy Caulfield. “Scientists’ perspectives on consent in the context of biobanking research.” *European Journal of Human Genetics* 23, no. 5 (2015): 569-574.

Mittelstadt, Brent Daniel, and Luciano Floridi. “The Ethics of Big Data: Current and Foreseeable Issues in Biomedical Contexts.” *Science and engineering ethics* (2015): 1-39.

Prainsack, Barbara, and Alena Buyx. “A solidarity-based approach to the governance of research biobanks.” *Medical Law Review* 21, no. 1 (2013): 71-91.

Rodriguez, Laura L., Lisa D. Brooks, Judith H. Greenberg, and Eric D. Green. “The complexities of genomic identifiability.” *Science* 339, no. 6117 (2013): 275-276.

Rothstein, Mark A. *Genetic secrets: protecting privacy and confidentiality in the genetic era*. Yale University Press, 1997.

Websites:

The Presidential Commission for the Study of Bioethical Issues. “Privacy and Progress in Whole Genome Sequencing.” http://bioethics.gov/sites/default/files/PrivacyProgress508_1.pdf

IGSR: The International Genome Sample Resource: http://www.internationalgenome.org

---

[1] ICD-9 codes stand for the “International Classification of Diseases, Ninth Revision” of the World Health Organization.

Notes

The author wishes to acknowledge the contributions of Karin Ellison, OEC - Life and Environmental Sciences Editor, and Joseph Herkert, OEC Engineering co-Editor. They provided valuable input in selecting topics and crafting the resources.

Contributor(s)

Valerie Racine
Karin Ellison
Joseph Herkert

Rights

Resource Type

Case Study / Scenario

Parent Collection

Big Data in the Life Sciences Collection

Topics

Big Data
Controversies
Data Management
Privacy and Surveillance

Discipline(s)

Genetics and Genomics
Life and Environmental Sciences
Research Ethics