Author: Valerie Racine
Created date: February 10, 2017
URL: http://www.onlineethics.org/40348.aspx
Note: 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.

Big Data & Genetic Privacy: Re-identification of Anonymized Data[1]

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,[2] 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.

[1] This material is based upon work supported by the National Science Foundation under Award No. 1355547, Karin Ellison and Joseph Herkert, Arizona State University sub-award Co-PIs. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

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

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?

Commentary

The study by Gymrek *et al.* 2013, and others like it, generated demands for additional restrictions in database sharing policies, changes to how and what kinds of data were collected and anonymized, and worries about some of the foundational concepts in research ethics, including the notions of informed consent, privacy, confidentiality, and the nature of the researcher/clinician – subject/patient relationship. This short commentary will focus on those concepts in biomedical research ethics.

Most researchers and ethicists agree that it is important to safeguard privacy and confidentiality for patients and research subjects, but to do so in a way that does not impede scientific progress. This "sweet spot" between the competing goals of scientific research and the individual's right to privacy is especially relevant for current genomic and genetic analyses using big data. For instance, Genome Wide Association Studies (GWAS) capitalize on correlated sets of large databases of individuals' genetic variants to determine whether certain variants are important contributors to complex diseases or disorders. There is also much optimism about the prospects of personalized medicine, in which medical professionals would access and integrate patients' personal genomic data into targeted and tailored treatments. The success of personalized

medicine, however, requires knowledge about which sorts of treatments will be effective for certain genetic variants, which depends on genomic analyses of big data.
While there are clear potential benefits of biomedical research analyses of large sets of genomic and genetic data, that information is also particularly sensitive as it can accurately reveal subjects' identity in the same way as social security numbers can. It can also reveal the identity of an individual's relatives. Because of the way this information can serve as accurate individual identifiers, some researchers have taken the notion of genetic privacy to denote a special instance of privacy (e.g. Rothstein 1997), based on the notion of "genetic exceptionalism" – i.e. "the view that being genetic makes information, traits, and properties qualitatively different and deserving of exceptional consideration" (Lunshof *et al.* 2008).

If we accept a concept of genetic privacy, based on genetic exceptionalism, then there are implications for the way we think about infringement of privacy and breach of confidentiality within the biomedical research context. For instance, Lunshof *et al.* (2008) argues that because some violations of privacy occur which are beyond the control of individuals or institutions (as in the above case scenario), they do not necessarily signal a moral failure even though those violations may cause harm in some instances. However, they note that the promise of confidentiality implies a relationship of trust and, with it, moral responsibilities on those who promise confidentiality. For that reason, a breach in confidentiality does entail a moral failure with respect to the relation of trust between the researcher/clinician and subject/patient.

These moral considerations have led research scientists and ethicists to rethink the model of informed consent that typically guides the relationships of trust between clinician/researcher and patient/subject in the biomedical context, and to reconsider what, if any, sense of privacy and anonymity should be promised to patients and research subjects.

Informed consent is typically used in cases of specific research studies. It is problematic in research that makes use of big data because it does not, and cannot, explicitly cover all future investigations, or future instances of sharing and aggregating data across research communities. Because of these elements in big data science, the traditional notion of informed consent cannot be implemented in the usual way.

Consequently, some have proposed more liberal notions of consent, such as "open," "broad," or "blanket" consent (Mittelstadt & Floridi 2015). These notions of consent require research participants to consent to all future research activities that makes use of their data. However, those approaches have been criticized for limiting patients' or subjects' autonomy (Mittelstadt & Floridi 2015; Master *et al.* 2014). An alternative proposal to the models of general consent is the notion of "tiered" consent. That notion of consent would enable patients and subjects to choose to limit future access to their data to only some kinds of research, or to require researchers to re-consent patients and subjects for specific kinds of future research. That approach has been criticized for creating too many difficulties for researchers and the management of large databanks.

Another alternative has been to emphasize the concept of solidarity rather than consent. This approach relies on the participation of "information altruists" concerned with the public good. It is mainly concerned with how research can be pursued and harms can be mitigated, "by providing data subjects with a ‘mission statement’, information on potential areas of research, future uses, risks and benefits, feedback procedures and the potential commercial value of the data, so as to establish a ‘‘contractual’’ rather than consent basis for the research relationship” (Mittelstadt & Floridi 2015; Prainsack and Buyx 2013). The proposed reliance on solidarity and public sentiment has been criticized for placing undue burdens on individuals to participate in research. However, it might also serve to emphasize the ethical responsibilities of big data researchers and database managers, and encourage scientists to be more proactive in the disclosure and transparency of risks of harm that might occur as a consequence of the loss of privacy (Lunshof *et al.* 2008; Barocas & Nissenbaum 2014). In this way, genomic and genetic research dependent on large sets of data has the potential to shift the moral responsibilities of researchers from protecting the privacy of individuals to ensuring the just distribution of any benefits from the outcomes of their research (Fairfield & Shtein 2014).

The emerging concepts of consent under negotiation within this research context, and the emphasis on researchers’ duty to benefit research participants and their communities more widely as well as the research participants’ duty to contribute to the public good, are areas of ethical deliberation intended to maintain the public’s trust in the medical profession, and scientific institutions more broadly. These ethical concepts and proposals, therefore, ought to be evaluated by how well they are able to do so.

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