Informed Consent and Privacy of De-Identified Information and Estimated Data

Lessons from Iceland and the United States in an Era of Computational Genomics

Donna M. Gitter
Department of Law
Baruch College, City University of New York
55 Lexington Avenue
New York, New York 10010
USA
e-mail: Donna.Gitter@baruch.cuny.edu

Abstract—**Advances in bioinformatics and computational genomics necessitate reexamination of the principles of privacy and informed consent. The law of informed consent requires that research subjects give their consent to participation in biomedical research. In the current age of bioinformatics and computational genomics, however, researchers are in many cases able to use genetic and genealogical data from research subjects who did agree to participate in genetic testing, in order to make educated guesses about the genetic profile of the subjects' relatives, who did not volunteer to participate. The law of informed consent does not address the use of estimated data, given that it was not possible before the advent of computational genomics to conduct "in silico" research. In considering whether to extend informed consent protection to those to whom "estimated data" is extrapolated, it is useful to consider currently proposed changes to the law of informed consent in the U.S. These proposed changes arise from the notion that biospecimens are increasingly considered intrinsically identifiable, and therefore individuals ought to be asked for their informed consent before the use of even de-identified specimens. Moreover, the recently revised Genomic Data Sharing (GDS) Policy of the U.S. National Institutes of Health (NIH) goes even further to require informed consent not only for use of biospecimens and identifiable private information, but also for genomic or other data, even if it is de-identified. It follows logically that those who do not agree to participate in biomedical research, but from whom estimated data are gleaned, ought to be asked for their informed consent.**

Keywords-bioinformatics; computational genomics; privacy; informed consent.

I. INTRODUCTION

Advances in bioinformatics and computational genomics necessitate reexamination of the principles of privacy and informed consent. Since the formation of the Nuremberg Code, which developed as a result of the Nazi War Crimes Tribunal and was the first internationally recognized code of research ethics, medical researchers must recognize protections for human research subjects. The primary tenet of the Nuremberg Code is that "The voluntary consent of the human subject is absolutely essential." In the current age of bioinformatics and computational genomics, however, researchers are in many cases able to use genetic and genealogical data from research subjects who did agree to participate in genetic testing, in order to make educated guesses about the genetic profile of the subjects' relatives, who did not volunteer to participate. This estimated data can then be combined with health records of the non-volunteers in order to conduct genetic research, often termed "in silico" biology, without their informed consent. Researchers use these technologies to calculate the probability that an individual carries a particular genetic variant, without sequencing that person's deoxyribonucleic acid (DNA), thereby developing estimated data for inclusion in research databases.

Section II of this paper considers the use of computational genomics in Iceland to conduct research using estimated data from individuals without their informed consent, noting that this conflicts directly with a legal trend toward enhanced recognition of the privacy rights and autonomy of research participants, as reflected in proposed changes to the law and policy of informed consent in the United States. Section III then considers proposed changes in the U.S. enhancing informed consent protection for research with de-identified materials, and advocates for the same level of protection for estimated data, in keeping with traditional norms of informed consent.

II. THE USE OF COMPUTATIONAL GENOMICS IN ICELAND

Controversial methods of computational genomics, particularly the use of estimated genetic data, are particularly effective in Iceland, an island nation with detailed genealogical records and a population of approximately 320,000 citizens who are considered to be genetically homogeneous. The intimacy of this small country is made evident by the existence of a smart-phone app in Iceland that permits individuals to determine whether they are related to another person whom they are considering dating.

 ISBN: 978-1-61208-457-2

In light of Iceland's genetic homogeneity and the availability of detailed genealogical information, in 1996 Icelander Dr. Kari Steffansson founded the company deCODE Genetics in order to use Iceland's population to pioneer genetic population studies. In 1999, the Icelandic government granted deCODE an exclusive 12-year license to build a Health Sector Database to hold centralized health records of its entire population [1]. The plan incited much controversy due to the presumption that citizens of Iceland would be deemed to consent to participate unless they actively opted out. In November 2003, the Supreme Court of Iceland disrupted deCODE's plans by ruling in favor of Ragnhildur Gudmundsdottir, an eighteen-year-old student, holding that she could prevent the transfer to the database of her deceased father's health records. The court held that the records in the database might allow her to be identified as an individual at risk of a heritable disease, even though the data would be anonymous and encrypted. The court noted that this risk was heightened by the fact that the Health Sector Database would allow information to be linked with data from other genetic and genealogical databases [1].

DeCODE then pursued another strategy, using estimated data to create a research database to find genetic sequences linked to diseases. Using DNA and clinical data from more than 120,000 research volunteers, deCODE analyzed their DNA sequences for a selection of slight variations called single nucleotide polymorphisms (SNPs), which are the most common genetic variations among individuals and some of which may prove important in the study of human health.

Using a relatively new technique, deCODE geneticists calculate the probability that an individual carries a particular genetic variant without actually sequencing that person's DNA. For example, deCODE was able to use its whole genome sequencing of the DNA of approximately 2,500 research participants in order to extrapolate the genomes of many more individuals. When deCODE identified a genetic variant of interest among the 2,500 whole genomes, the company used the more limited SNP data that it had amassed from its 120,000 volunteers in order to impute, with 99 per cent accuracy, whether any among the 120,000 also carried the mutations [6]. As noted by one source, "if your mother had been in the hospital for a stroke and agreed to participate in a clinical study, while her brother had volunteered his DNA, deCODE would be able to predict *your* likelihood of a genetic disposition for stroke [5]."

While other researchers are using the same technique as deCODE, the company's unique approach is to combine the known and estimated genotypes for its research participants with its genealogical database, thereby permitting deCODE to estimate what it calls the "in silico" genotypes of close relatives of the volunteers whose SNPs were analyzed. This permits deCODE to infer data about 200,000 living and 80,000 deceased Icelanders, who have not consented to participate in deCODE's studies. Further, it could give the company genotypes for the largely consanguineous population of 320,000 people in its entirety. Researchers can then determine whether a variant in a DNA sequence found by fully sequencing the DNA of a small group likewise appears in a larger population in the same proportion [6].

The company has used these estimated genotypes for individuals as controls in its studies and also combined them with health records for patients who were involved in a disease study in Iceland but whose DNA has not been sampled. Using estimated data, deCODE published six papers between 2011 and 2013 in the prestigious journals *Nature*, *Nature Genetics*, and the *New England Journal of Medicine*, linking specific genetic mutations to risks of diseases. DeCODE's drug discovery efforts were less successful, however, and the company declared bankruptcy in 2009. In December 2012, Amgen purchased the company for $415 million [6].

In 2012, deCODE planned to use its strategy as part of a new study. Having imputed the genotypes of the close relatives of the volunteers whose SNPs had been fully catalogued, deCODE intended to collaborate with Iceland's National Hospital to link these relatives to certain hospital records for individuals, such as surgery codes and prescriptions. On May 28. 2013, Iceland's Data Protection Authority (DPA) denied this request, on the grounds that it would violate the relatives' privacy unless they gave their informed consent. The DPA gave deCODE until November 2013 to demonstrate that it obtained consent [10].

DeCODE ultimately found a means of working around the requirement of informed consent, describing it in a November 5, 2013 letter to the DPA. DeCODE confirmed that it had deleted all data registers containing imputed genotypes for individuals from whom consent was lacking. However, deCODE also presented the DPA with a proposal, according to which genotype data from research participants (who had consented) would be linked with genealogy data in a way that would generate statistical results as strong as those formerly achieved. According to the Iceland DPA, this would entail that a genetic imputation for those who had not consented would be generated "in a split [] second in the processing memory of a computer. However, this imputation would then cease to exist and would never be accessible to anyone in any form. The only accessible data would be the aforementioned statistical results, which would not in any way be traceable to individuals [10]." The DPA confirmed in a letter dated 26 November 2013 that this proposal did not give rise to objections if "all the aforementioned prerequisites were met [10]."

Most recently, deCODE published a series of papers in the journal *Nature Genetics* in March 2015 that described sequencing the genomes of 2,636 Icelanders, the largest collection ever analyzed in a single human population. Using the imputation technique, deCODE claims that it was able to combine the full genomes it has for about 10,000 Icelanders and the partial genetic information on 150,000

 ISBN: 978-1-61208-457-2

more to generate a report for genetic disease on every person in Iceland. For example, the firm can identify every Icelander with the well-known BRCA2 mutation, which raises the risk of breast and ovarian cancer, even if the individuals have not submitted to genetic testing themselves.

Dr. Steffánsson of deCODE contends that his company's research methods do not violate patient privacy because the company is not actually sequencing the citizens' DNA, but rather devising "conjectures" or "hypotheses" about them, rather than obtaining personal information. He notes that estimated DNA sequences, unlike directly measured sequences, are not very accurate for individuals, though they are valuable at the group level. Moreover, Steffánsson emphasizes that, until now, both the DPA and Iceland's national bioethics committee have approved the use of estimated genotypes for the two-thirds of Icelanders who have not consented to its research [6].

Geneticists disagree as to whether deCODE must obtain informed consent. Jón Jóhannes Jónsson, a geneticist with the University of Iceland, observes that deCODE is not truly doing anything new, given that geneticists routinely infer whether relatives who are not part of a particular study carry a genetic mutation. What is different about deCODE's strategy is that it invokes the DNA sequences of the entire Icelandic population. Jónsson concedes that deCODE's plan to use estimated data supplemented by hospital records presents a difficult case. Daniel MacArthur, a geneticist at Massachusetts General Hospital in the United States, suggests that although deCODE did not actually violate the privacy of individuals, from an ethics points of view the researchers should at least attempt to obtain informed consent. MacArthur laments that blocking deCODE from using its estimated data present a "tragedy" not only for the company, but the wider "complex disease genetics community [6]."

On the other hand, DeCODE's promise to delete individuals' data once it has calculated statistical results remains problematic, given the increasing proliferation of easy, cheap, and powerful reidentification technologies. [8] Erlich and Narayanan, experts in computational biology and computer information systems, have deemed deCODE's actions a "breach" of "genetic privacy" of the sort increasingly common in the last few years as the range of techniques to carry out such privacy breaching "attacks" has expanded. In particular, they term deCODE's method a "completion technique," meaning the use of known DNA data "to enable prediction of genomic information when there is no access to the DNA of the target." There have been several high profile breaches of privacy whereby an "attacker" has been able to infer, from the known genome of one individual, the genomes of his or her relatives [3].

Erlich and Narayanan note that deCODE's approach is an advanced version of the completion technique, given that deCODE has access to the genealogical and genetic information of several relatives of the target, and permits genotypes of distant relatives to be inferred. They explain that it is possible to develop an algorithm that finds relatives of a "target" who donated their DNA to the reference panel and who share a "unique genealogical path that includes the target, for example, a pair of half-first cousins when the target is their grandfather [3]." A shared DNA segment between the relatives indicates that the target has the same segment. By studying more pairs of relatives that are connected through the target, it is possible to collect more genomic information on the target without any access to his or her DNA, and, more importantly, without his or her informed consent [3]. This conflicts directly with a legal trend toward enhanced recognition of the privacy rights and autonomy of research participants, as reflected in proposed changes to the law and policy of informed consent in the United States.

III. PROPOSED CHANGES TO TO THE LAW AND POLICY OF INFORMED CONSENT IN THE U.S.

The September 8, 2015 Notice of Proposed Rulemaking (NPRM) published by the U.S. Department of Health and Human Services in the Federal Register, entitled *Human Subjects Research Protections: Enhancing Protections for Research Subjects and Reducing Burden, Delay, and Ambiguity for Researchers*, reflects the emerging recognition of the dangers of re-identification of research participants [4].

In the summary of its major provisions, the NPRM provides that "informed consent would generally be required for secondary research with a biospecimen (for example, part of a blood sample that is left over after being drawn for clinical purposes), even if the investigator is not being given information that would enable him or her to identify whose biospecimen it is [4]." The NPRM describes the changes in technology driving this proposed change, noting that "[n]ew methods, more powerful computers, and easy access to large administrative datasets produced by local, state, and federal governments have meant that some types of data that formerly were treated as non-identified can now be re-identified through combining large amounts of information from multiple sources," including publicly available sources. In light of this change, "the possibility of fully identifying biospecimens and some types of data from which direct identifiers had been stripped or [which] did not originally include direct identifiers has grown, requiring vigilance to ensure that such research be subject to appropriate oversight [4]." "Most importantly", according to the NPRM, "[a] growing body of survey data shows that many prospective participants want to be asked for their consent before their biospecimens are used in research [4]." Thus, the NPRM clearly prioritizes an individual's right to elect or decline participation in research. This notion aligns with recognition of the right of informed consent for individuals who participate via in silico biology, though the use of their estimated data.

 ISBN: 978-1-61208-457-2

Moreover, the U.S. National Institutes of Health (NIH) have recently revised their Genomic Data Sharing Policy (GDS) to set forth the expectation that investigators will obtain participants' consent not only for the use of their biospecimens and identifiable private information, but also for the use of their genomic data. This will be true even if the cell lines or clinical specimens used to generate the data are de-identified [7]. By requiring informed consent for genomic data, the GDS goes even further than the NPRM is recognizing the risks of re-identification and an individual's right to informed consent for research participation.

There are many reasons that individuals may object to the use of their de-identified information, even if it is estimated data. First, individuals may decline on ethical, religious or other personal grounds to participate in certain controversial forms of research, such as somatic nuclear cell transfer, stem cell research, and germ-line gene therapy. As noted in the Human Subjects Research NPRM, "a more participatory research model is emerging in social, behavioral, and biomedical research, one in which potential research subjects and communities express their views about the value and acceptability of research studies [4]." Second, research participants may object to commercial exploitation of discoveries developed through the use of their de-identified information. Largely in response to some highly publicized lawsuits in which research participants have sued researchers for revenue earned from using their information and biospecimens, it has become common for researchers to present research participants with informed consent documents that disclaim any economic interest in possible commercial applications flowing from the research. Research using de-identified records is highly problematic in that there is no informed consent and therefore no disclaimer.

Just as there are many valid arguments in favor of expanding informed consent protections for research participants, there are numerous reasons why the research community is likely to oppose the extension of research protections, whether for de-identified biospecimens or information, or estimated data. First, it is not feasible to contact each individual from whom materials have been gathered in order to request that person's informed consent. Even if it were possible, it would be very time-consuming and costly. Each individual's contribution to the research is so small, perhaps as to be dispensable, yet would require the full process of informed consent. Most importantly, and flowing from these reasons, the necessity of such informed consent might delay and perhaps even preclude altogether the development and introduction of medical advances. Furthermore, it is not only researchers, but also patient advocacy groups, who warn of these dangers. As noted by these critics, in the context of requiring informed consent for the use of de-identified biospecimens and identifiable private information, requiring such consent "might inappropriately give greater weight to the [] principle of autonomy over the principle of justice, because requiring consent could result in lower participation rates in research by minority groups and marginalized members of society," though "most of the comments from individual members of the public strongly supported consent requirements for use of their biospecimens, regardless of identifiability [4]."

Indeed, it can undermine trust in the medical establishment when individuals learn that their biospecimens or information, whether de-identified or estimated, are used without their consent. Indeed, the Human Subjects Research NPRM states that "the failure to acknowledge and give appropriate weight to this distinct autonomy interest in research using biospecimens could, in the end, diminish public support for such research, and ultimately jeopardize our ability to be able to conduct the appropriate amount of future research with biospecimens [4]."

It is clear that the trend, as evidenced by the Human Subjects Research NPRM and the revised NIH GDS, is toward the requirement of informed consent for the use of de-identified biospecimens and genetic information. The question then arises whether there is a meaningful distinction between de-identified biospecimens and information, on the one hand, and estimated data, on the other, in terms of the need for informed consent. It should be noted that neither de-identified information nor estimated data requires any direct interaction with the individual about whom it is gathered. Indeed, the Common Rule specifies that human subject research occurs when an investigator conducting research obtains "data through intervention or interaction with the individual", or obtains "identifiable private information" from any source [2]. The regulation further provides that "Private information must be individually identifiable (i.e., the identity of the subject is or may readily be ascertained by the investigator or associated with the information) in order for obtaining the information to constitute research involving human subjects [2]." It is this condition of individual identifiability that deCODE Genetics seeks to avoid when it declares to the Icelandic Data Protection Authority that the data will be individually identifiable only for a split second and then deleted from the computer memory. This argument fails, however, if data are as easily identifiable as Yaniv and Erlich have described.

The main difference between de-identified biospecimens and identifiable private information, on the one hand, and estimated data, on the other, is that the latter are not accurate at the individual level, but only at the group level**.** While this fact may adequately address the privacy issue, it does not resolve the issue of autonomy, meaning individuals' ability to decline to participate in research, either totally or as a means of rejecting the specific research proposed.

IV. CONCLUSION AND FUTURE WORK

Biospecimens are increasingly viewed as intrinsically identifiable. What is more, armed with bioinformatics and computational genomics techniques, along with public and private databases, researchers can accurately impute the genetic sequence information of individuals without their informed consent. While this can yield new discoveries and vital data for improving diagnostics, it also raises complex questions regarding the need to obtain informed consent from research participants about whom data is imputed via

in silico research. The law of informed consent, codified before the development of powerful current technologies, does not address issues arising from the use of estimated data.

Proposed changes to U.S. regulations would provide enhanced protection for research subjects by requiring informed consent for the use of their biospecimens and identifiable private information, whether clinical or from prior research. Presently, researchers can use these specimens without consent by stripping them of identifiers. The newly revised NIH GDS goes even further by requiring informed consent for the use of genomic data, even if it derives from de-identified sources. These changes reflect the current view that researchers ought to respect the privacy and autonomy of research participants in an era where re-identification of research subjects has become easier to achieve. While a liberal reading of the proposed federal rule changes and the new NIH policy support the notion that those from whom estimated data is gathered and used are entitled to the same rights of informed consent, privacy, and autonomy as conventional research subjects, the proposed rule changes contemplate for the moment only research subjects who contribute biospecimens or identifiable private information, whether wittingly or not. This article contends that individuals who contribute estimated data are similarly entitled to be asked for their informed consent for their research participation.

The next steps in this research will be an investigation of the "right not to know" the results of one's genetic risks. Paradoxically, while the law provides increasing protection for the right of informed consent, there is an emerging view that genetic incidental findings ought to be gathered and returned to individuals, even absent their informed consent. Indeed, deCode declares that it ought to be able to contact Icelanders to inform them of the genetic risks of which deCode learned when studying their estimated data. This raises the troubling specter of individuals who have given consent neither for the use of their estimated data, nor the return of incidental findings to them, having their data used for research and then being contacted with researchers' incidental findings. This paternalistic approach conflicts deeply with the longstanding norms of biomedical ethics.

ACKNOWLEDGMENT

The author thanks for their helpful comments on this work the hosts of and participants at the Law and Ethics of Big Data Research Colloquium held at University of Indiana at Bloomington on April 18, 2015. The author also expresses appreciation to the Zicklin School of Business, Baruch College, City University of New York, for research support for this work.

REFERENCES

[1] A. Abbott, "Icelandic database shelved as court judges privacy in peril," Nature, vol. 429, p. 118, May 13, 2004, doi:10.1038/429118b.

[2] Code of Federal Regulations, 45 C.F.R. § 46.102(f) (2009).

[3] Y. Erlich and A. Narayanan, "Routes for breaching and protecting genetic privacy," Nature Reviews Genetics, vol. 15, pp. 409-421, May 8, 2014, doi:10.1038/nrg323.

[4] Federal Register, Federal Policy for the Protections of Human Subjects: Proposed Rules, Vol. 80, No. 172, 53,933 – 54,060 (September 8, 2015).

[5] R. Goldin, "Privacy and our genes: is deCODE's DNA project 'Big Brother' or the gateway to a healthier future," Genetic Literacy Project, June 24, 2013, available at https://www.geneticliteracyproject.org/2013/06/24/privacy-and-our-genes-is-decodes-dna-project-big-brother-or-the-gateway-to-a-healthier-future/, retrieved: January, 2016.

[6] J. Kaiser, "Agency nixes deCODE's new data-mining plan," Science, vol. 340, pp. 1388-1389, June 21, 2013, doi: 10.1126/science.340.6139.1388.

[7] National Institutes of Health, NIH Genomic Data Sharing Policy, Notice Number NOT-OD-14-124 (Aug. 27, 2014), http://grants.nih.gov/grants/guide/notice-files/NOT-OD-14-124.html, retrieved: January, 2016.

[8] P. Ohm, "Broken promises of privacy: responding to the surprising failure of anonymization," UCLA Law Review, vol. 57, pp. 1701-1777, 2010.

[9] T. Sveinsson, Iceland Data Protection Authority, E-mail message to Donna M. Gitter, Professor of Law, Baruch College (Oct. 20, 2014), unpublished.

[10] K. Yandell, "All Icelandic women with the BRCA2 gene can be found in the database," News of Iceland, May 13, 2013.

 ISBN: 978-1-61208-457-2