Published in final edited form as:
Soc Res (New York). 2011 ; 78(3): 907–932.

Bodies of Data: Genomic Data and Bioscience Data Sharing

Pilar N. Ossorio

THE LIFE SCIENCES HAVE UNDERGONE A RADICAL SHIFT FROM small-scale, single-molecule, laboratory-based research to large-scale, *in silica* research, in which tens of thousands of genes, transcripts, and/or proteins can be studied simultaneously (Zweiger 2001: xi). Genomics has become the model for large-scale biology. Human genomics is large-scale in terms of the numbers of people whose biological materials are used, specimens and nucleotides studied, and researchers involved in each project. The science also produces enormous quantities of data. Over the past decade, genome research data have increased in complexity, heterogeneity, resolution, and volume (Ostrom 2006: 341).

In the contemporary research environment, data from genomics projects often do not stay in the laboratory in which they were generated. Research funders in the United States and United Kingdom have increasingly imposed data-sharing requirements on their funded scientists. A variety of policies imposed by the National Institutes of Health (NIH) in the United States, the Welcome Trust in the United Kingdom, and other funders require that researchers deposit data in repositories that are either entirely open to the public or that make data available to a more limited set of applicants—qualified scientists with appropriate research projects. The latter mechanism is referred to as "controlled access."

Funders require data sharing because it may leverage the public's enormous investment in genome research by promoting studies that combine datasets, allowing different analytical methods to be used on a single dataset, encouraging different hypotheses to be tested, and allowing widespread use of unique or extremely expensive data. Nonetheless, widespread dissemination of vast quantities of individual-level data derived from humans raises concerns about respect and harm to persons. Controlled-access mechanisms are designed to protect "data sources" from informational harm while allowing the widest possible dissemination of data. An important question is whether controlled access provides adequate protections.

Scholars have generated a robust literature on biobanking, in which they explore the ethical, legal, and social implications (ELSI) of the storage and use of large quantities of extracorporeal human biological materials. Far less attention has been paid to the implications of storing and sharing *data* pertaining to individual human beings whose biological materials and personal information are used in research. I argue that scholars ought to address questions raised by the generation and sharing of large quantities of individual-level, biomedical data separately from concerns about the storage and distribution of biological specimens, because the two types of resources are subject to different legal regimes, stored in different types of institutions, and subject to different professional norms and networks. I then describe one major policy and institutional infrastructure for sharing controlled-access data. Finally, I discuss some results from a pilot empirical study designed to investigate whether controlled-access policies are likely to prevent unauthorized disclosure or inappropriate uses of the personal information and genome data being shared.[1]

Today there has been a radical transformation of *biology* into an *information science* (Ostrom and Hess 2006: 335; emphasis in original).

Throughout this paper, my claims and arguments draw on this pilot study and on my experience as an adviser to several large genome research projects. A word about nomenclature is also in order at this point. There is no elegant, succinct terminology for describing people from or about whom genomic data have been derived. “Research participant” is not the appropriate terminology for two reasons. First, some genomic data are generated without the knowledge or consent of the people to whom they pertain. This can happen if, for instance, extrinsic identifiers are stripped from tissue and medical information that are then used in genome research. Alternatively, an institutional review board (IRB) may grants waivers of consent and authorization for research using existing biological specimens and medical information. The term “participant” is meant to connote human research subjects’ active engagement and agency, which are absent in studies done without their knowledge or consent.

Second, some scholars and regulatory professionals use the phrase “research participant” in lieu of “human subject,” a regulatory term with precise legal and technical meanings that do not apply to all cases in which scientists generate and use genomic data and other personal information. Thus, neither “human subject” nor “human participant” is a globally appropriate term, and I only use these designations when federal regulations for the protections of humans in research would apply.[2]

Having ruled out “research participant” and “human subject,” I have reluctantly chosen to use the term “data source.” In some respects, the remote and abstract connotations of “data source” are appropriate for a study of scientists who use controlled-access data. Most of these scientists have never met or interacted with the human individuals to whom the data pertain. Large-scale science has meant increasing specialization, in which some scientists (usually physicians, nurses, or allied health professionals) collect specimens and personal information while other scientists generate data, and yet others manipulate and analyze data. Scientists who obtain data from repositories—data users—are often PhDs in the biological or computational sciences. Data users may have little direct experience with or interest in a medical condition represented in their data. A substantial gulf—social, emotional, and geographic—often separates the data-user from her sources, and thus “data source,” an impersonal term that carries a sense of emotional distance, may be both technically accurate and connotatively appropriate.

DATA AND THE DISEMBODIED PERSON

Inputs to genome research include DNA or RNA culled from spit, blood, stool, or biopsies. From this material the research generates various forms of genetic and other molecular data. Personal and medical information—acquired from medical records, questionnaires, interviews, and physical exams—constitutes another input to genome research. For example, the US-funded Human Microbiome Project (Turnbaugh 2007) collects detailed descriptions of participants’ sexual and hygiene practices, medication histories, personal and family medical histories, locations where each participant has resided, smoking behavior, employment history, educational information, and income. Datasets available for sharing

[1]The study involved semi-structured, in-person interviews with 17 users of controlled-access data (data users) and 3 members of committees that oversee the release of such data. The interviews were transcribed, coded, and analyzed using inductive, Grounded Theory methods (Strauss 1987). Interview subjects were chosen from publicly available lists of scientists who have accessed controlled data, and from lists of published journal articles in which controlled-access datasets have been used. Interviewees included scientists from the United States, the United Kingdom, and China. This project was reviewed by the University of Wisconsin IRB and determined to be exempt from the requirements of the federal Common Rule (DHHS 2005).

[2]The relevant regulations are the federal Common Rule, 45 CFR §46 (2005), and the comparable FDA regulations, 29 CFR §§ 50 and 56 (2009).

often contain large quantities of highly complex and personal information in addition to genome data.

Genome data include whole-genome genotypes, whole-genome sequences, or whole-exome sequences.[3] The science aims to generate information about the relationship of genetic markers[4] in the sequences or genotypes to traits, health outcomes, or biological states, and to provide the initial information from which scientists can develop knowledge of the biological pathways and mechanisms that produce disease or other states of interest. Some very large projects aim to create "resources" of information, analytical tools, and biological materials for entire fields of scientists. The Human Genome Project (HGP) was a resource-generating project.

The quantity and complexity of today's genome data rival that produced by physicists. By 2010, the 1000 Genomes Project had generated 4.9 terabases of DNA sequence for its Pilot Phase; the Cancer Genome Atlas was producing approximately 7.3 terabases of sequence per month (Consortium 2010; Ozenberger 2011). (The prefix "tera-" in terabase or terabyte means 10^{12} or 1 trillion.) By mid-2010, researchers at the Wellcome Trust were sequencing more DNA every two seconds than was sequenced during the entire first five years of the HGP (Berger 2010). Knowledge production in the life sciences involves the ability to manipulate and analyze these data; "biology is being reborn as an information science" (Zweiger 2001: 17).

ELSI scholars have focused predominately on what happens to, and how people feel about, the material inputs to genome research. The robust literature on biobanks focuses squarely on the "bio," with little regard for the differences in the legal regimes, institutional forms, or professional norms applicable to the concomitantly collected information or the data generated. Yet there are many differences between biological materials and data that render the latter eminently worth studying in their own right. Specimens and data require different skills to properly describe, standardize, store, and curate. They are often stored at different institutions, which are governed by different formal law, institutional policies, and informal norms. Data and biological material may be accessed and used by different people.

In large genomics projects and, increasingly, in many smaller ones, the specimens are sent by a "collecting researcher" to a specimen repository. Such repositories are run by people with training in biology, pathology, and other medical disciplines. Repositories for biological materials are composed of wet laboratories and possess the freezing and storage capacity to hold large numbers of physical specimens. Repository personnel spend their time developing media with which to nourish cells, creating cell lines from blood or other specimens, and extracting DNA from cells. Their quality control activities include measuring cell survival rates and analyzing specimens for biological contaminants.

Data and tissue often reside in physically, legally, and professionally different institutions. In contrast to tissue repositories, data repositories are developed and managed by people with professional expertise in computing and bioinformaticians. Personnel at data repositories spend their time designing computational "pipelines" through which data generators can upload data to the repository and users can download it. Quality control

[3]A genotype is an ordered catalog or map of the particular DNA variants a person possesses. For a large project, each person's genotype would typically consist of data from about 500,000 to 1 million genetic markers spread relatively evenly across a person's entire genome. "Whole-exome" sequence refers to sequence representing all of the protein coding regions in a person's genome. Only a few percent of the human genome encodes proteins.

[4]A "marker" is any place where the DNA varies between people. This variation can be a single nucleotide polymorphism (SNP), a site in the genome where some people have a particular nucleotide (one of the four chemical building blocks of DNA), such as guanine, and other people have a different nucleotide, such as cytosine. Other types of markers involve different numbers of repeated motifs that people may have at a particular site, or insertions or deletions of sequence.

activities at data repositories include ensuring that data elements are not repeated, that inappropriate data are not uploaded, and that the data are properly annotated. Repository personnel help to design controlled vocabularies for data elements, determine which elements will be included in the repository, devise standards for data and metadata, and devise display options.

Repositories for data and biological material are also subject to overlapping but not identical legal regimes.[5] Some nations have laws specifically governing biorepositories; in the United States, state public health law generally governs the handling and proper disposal of specimens, but those laws do not apply to data. On the other hand, data and information may be subject to privacy laws that do not apply to biological material. For instance, the Standards for Privacy of Individually Identifiable Health Information of the Health Insurance Portability and Accountability Act (HIPAA Privacy Rule) (DHHS 2002) do not apply to research tissues that have been stripped of extrinsic identifiers, but they may apply to similarly stripped medical information or research data, if those data are held or used by covered entities (DHHS 2004). Europe has blanket data privacy laws that apply to biomedical data in repositories but not to tissues. Data and information may be intellectual property—trade secrets, patents, or copyrights; whereas, biological material is a form of personal property.[6]

The biological inputs to genomics are (mostly) rivalrous resources. Although we speak of "immortalized cell lines," many (perhaps most) cells cannot replicate forever. Tissue can be used up. And while DNA can be cloned, doing so extracts it from its usual biological context. In many cases, the biological material, is finite, so scientists' relations to people who give tissue for research, to the material itself, and to each other in the context of this material may be suited to analysis under ethical, economic, and political theories pertaining to resource competition and scarcity.

Data, on the other hand, are more in the nature of a nonrivalrous resource. Once a dataset is in a repository it can be copied, downloaded, and analyzed by an unlimited number of researchers. One person's use does not diminish the resource for others, and data are not consumed through use: they are not "used up." Data can be distributed to far more people than can specimens. The limits on data distribution are the limits of a nation's and an institution's broadband and data storage capabilities. Ethical, economic, and political theories of common goods and "club resources" are, perhaps, better suited to analyses of genome data (Contreras 2011; Ostrom 2006).

GENOME SCIENCE AND INFORMATIONAL HARM

Most risks of genomics research are informational—risks of having information used to the detriment of the data source. A person who contributes specimens for sequencing, genotyping, or other molecular analysis in research is not at risk of being overdosed, having a bad drug reaction, or having an experimental medical device malfunction. Instead, she risks having personal information from her genome, medical history, or life history used in ways that violate her rights or set back her psychological, social, economic, or legal interests.

[5]Because both types of repositories are institutions, both will be subject to the same general business law, such as employment, labor, and tax law.

[6]Cases such as *Moore v. Regents*, 973 P. 2d 479 (CA 1990), *Greenberg v. Miami Children's Hospital*, 264 F. Supp. 2d 1064 (S.D. FL 2003), and *Washington v. Catalona*, 490 F. 3d 667 (CA 8, 2007) may cause some to doubt that courts will treat extracorporeal tissue as property of the persons from whom it was derived, but there is no doubt that courts have treated the tissue as researchers' or research institutions' personal property.

Traditionally, anonymity was the governance tool that regulatory regimes and research ethics used to minimize informational risks to research participants. Protections contained in the widest reaching United States federal research regulation—the Common Rule—turn on whether living persons, or their data, are *individually identifiable* to researchers (DHHS 2005). If specimens or data are deemed individually identifiable then research using them is regulated as research on human participants. De-identified, or anonymized, specimens or data are viewed as posing little risk and, therefore, as needing little or no regulatory oversight. Anonymity ensures that a person cannot be named, located, emailed or otherwise become known and linked to sensitive information, because nothing in the research record points back to her.

The conventional means of de-identifying, or anonymizing, specimens and data has been to strip them of extrinsic identifiers, such as names, medical record numbers, or Social Security numbers. This approach may still work for specimens, but it no longer works for the data generated from them.

The traditional distinction between “identifiers” and other types of information about a person reflects a particular conception of personal identity in which identity is separable from the data source’s attributes (Agre 1997). This distinction breaks down, however, in the face of statistical methods for individuating and identifying data sources based on the accumulation of numerous, seemingly innocuous, pieces of data (Ohm 2010). With ever increasing computing power in the hands of more and more people, with the ability to combine databases from numerous sources, and with algorithms that can combine elements both within and across datasets, information scientists recognize that almost any datum could be identifying in some contexts or combinations.

Three aspects of contemporary genomic databases ought to raise the specter of informational risks to data subjects. First, the type of individual-level genomic data in these databases is intrinsically individuating and, by itself, could serve as an excellent identifier (Lowrance 2007; McGuire 2006). Second, many human genome research databases contain large quantities of sensitive medical and other personal information that might be combined to re-identify a data source. Third, some data elements in genome research databases (including genome data) might be similar or identical to elements in nonresearch databases, and might be used to link research databases with commercial, forensic, or other databases. Such linking can lead to the identification of a data subject and to the creation of a much more detailed profile of the data source than was originally available to either the researcher or the compiler of the other database.

If a source’s genome data is intrinsically identifying, then any medical or other personal information associated with the genome data also cannot be effectively anonymized. In 2007 the NIH acknowledged this state of affairs in introductory remarks to its published data sharing policy: “the NIH takes the position that technologies available within the public domain today, and technological advances expected over the next few years, make the identification of specific individuals from raw genotype-phenotype data feasible and increasingly straight-forward” (Zerhouni 2007). Therefore, the agency reasons, releasing controlled-access data in response to a request under the federal Freedom of Information Act (FOIA) would constitute an unreasonable invasion of privacy. The agency foresees that, in response to a FOIA request, it would redact all individual-level genotype and phenotype data.

The intrinsically identifying nature of genome data, and potentially, of personal information in genome research datasets, disrupts existing regulatory trade-offs and practices. IRBs and researchers are confused about whether to treat studies that use data downloaded from

genome repositories as research on human participants. IRBs are reluctant to require review and approval of research using data from which extrinsic identifiers have been removed, because such data have been de-identified according to traditional criteria and no longer fall under the IRB's purview. Yet, the quote above makes clear that NIH recognizes contemporary genome data, and any accompanying medical information, as intrinsically capable of identifying somebody. There is no regulatory category for such data. Ambiguity about the nature of genome data as a regulatory object, along with concern for the welfare of data sources, has led NIH and other research funders to develop new governance mechanisms for data repositories.

CONTROLLING ACCESS TO DATA

There are quite a few repositories that serve data (and even research results) to promote sharing among scientists. For genome data, the largest repository with the most formalized governance policy is the database of Phenotypes and Genotypes (known as dbGaP) run by the National Center for Biotechnology Information. dbGaP was created to instantiate the requirements of the "Policy for Sharing of Data Obtained in NIH Supported or Conducted Genome-wide Association Studies (GWAS)," which was published in the *Federal Register* in August 2007 (Zerhouni 2007). Formally, the policy only applies to GWAS; it says nothing about sharing DNA sequence data. However, various NIH institutes have used their funding power to require investigators whose NIH-funded research produces sequence data to release them through dbGaP, according to the GWAS policy. Other repositories, such as the Wellcome Trust Case-Control Consortium repository, operate according to similar policies, which I do not describe here for reasons of brevity and because the GWAS policy is the best documented and most widely used one.

The GWAS policy requires data producers to submit "descriptive information about their studies for inclusion in an open access portion of the database" (Zerhouni 2007: 49295). "Open access" means that the information is available on an unrestricted website. Descriptive information includes "protocols, questionnaires, study manuals, variables measured, and other supporting documentation," such as consent forms (49295). Some summary or aggregate results may be publicly released. Since 2008, however, when a group of researchers demonstrated that it was possible to pick an individual's genotype out of aggregate data, no aggregate results have been publicly released (Homer 2008).

The policy also "strongly encourages the submission of curated and coded phenotype, exposure, genotype and pedigree data, as appropriate" (Zerhouni 2007: 49295). In practice, any type of personal information that is collected about a data source can and will be deposited in dbGaP.[7] The repository will make these detailed, *individual-level*, raw or lightly computed data available to qualified scientists—data users—through a controlled-access procedure. Although the policy does not require the submission of individual-level data, my interviews indicated that NIH officials and scientists all expect that genotype and phenotype data from all or nearly all newly funded genome-wide studies will be deposited. And because each NIH institute can make its own, additional rules for determining which data should be deposited, when they should be deposited, and under what conditions they should be released, the GWAS policy is only a starting point for understanding the formal rules and informal norms governing data sharing through dbGaP.

[7]Importantly, this means that the Genetic Information Nondiscrimination Act of 2008, Public Law No. 110-233, 122 Stat. 881, does not apply to all of the individual-level information in dbGaP. Employment or health insurance discrimination could occur based on the nongenetic information in the repository.

The GWAS policy outlines the following mechanism for data submission: a submitting scientist and an appropriate official from her institution must certify that submission of the data is consistent with applicable laws, regulations and institutional policies (Zerhouni 2007: 49295). The submitting institution must describe, in detail, uses of the data that are consistent with or precluded by the original informed consent, if there was any consent. The submitting institution must also take steps to anonymize the genotype and accompanying information. It must strip the data of a variety of specified extrinsic identifiers and must not submit any data that it actually knows could be used, alone or in combination, to re-identify individuals. For coded data, the submitter must not transmit decryption keys to dbGaP, and must certify that it will not otherwise disclose research participants' identities to NIH. This last requirement seems devised to ensure that dbGaP cannot engage in activities that would count as regulated research under the Common Rule.

After the repository accepts a project's individual-level data, a data use committee oversees requests by prospective users to withdraw data from the repository. Researchers who request data must submit a brief protocol describing how they plan to use the data. Their plans must be consistent with any restrictions described by the data submitter or their request will be refused. A data requester and an appropriate official from her institution must also sign a contract (a Data Use Certificate) stating that the investigator will not attempt to identify any individual in the data set, will not transmit the data to anybody who is not named in her data access request, will use the data only for the approved research, will follow specified data security procedures and applicable laws, and will report herself for any violations of the policy (Zerhouni 2007: 49296). A few researchers have reported themselves for violations and have been told they had to stop using the dataset(s) at issue until the violation could be rectified.

In reviewing this policy one can identify significant inconsistencies regarding the identifiability of individual-level data in the controlled-access portion of the repository. These inconsistencies reflect the tension between NIH's desire to avoid having the repository come under the jurisdiction of the Common Rule and its attempts to protect data sources' interests. Essentially, under the policy the data are deemed anonymous enough that the Common Rule does not apply, but identifying enough that they would not be released in response to a FOIA request.

Because identifiability and anonymity are two ends of a spectrum, and real-world data often fall somewhere in the middle, logic does not preclude the possibility that data could be anonymized for some purposes and not others. However, both the Common Rule (as applied to data) and the FOIA exception are supposed to protect data sources from the harms that could materialize if sensitive information is associated with them. Given that these two legal rules have similar aims, NIH's position appears incoherent. A possible justification for treating identifiability differently under FOIA is that some people might make FOIA requests with the express purpose of trying to identify data sources, whereas federally funded scientists have other reasons for using the data and disincentives for identifying data sources. Thus, one might be justified in making different assumptions about the probability of both re-identification of and harm to data sources in the research context versus the nonresearch contexts in which a dataset might be "FOIA'ed."

Interestingly, although genome scientists often describe the Common Rule with bitterness, and feel that it restricts their research too severely in relation to the research risks, they and their agency have created a data governance system that recapitulates many features of the Common Rule's IRB-based system. The GWAS policy (and similar policies that govern other genome data repositories) requires protocol submission, and review of that protocol by a committee that can approve, defer, or deny access to a dataset. The data access committee

is quite similar to an IRB. However, the GWAS policy also differs from the Common Rule in some important ways. Notably, the GWAS policy attempts to make the promises of investigators who collected the specimens "run with" the data. The policy uses contract law in the attempt to bind researchers who withdraw data from the repository (and their institutions) to the promises originally made to actual research partici-pants and to other provisions intended to protect data sources. Even though dbGaP's data users do not interact with data sources, the users are contractually bound to behave in ways consistent with sources' interests. Thus, promises run with the data because the promises made by the researcher who initially collected the data become, in effect, legally attached to the data and bind subsequent users.

If the GWAS policy were universally and strictly observed, risks to data sources would be substantially minimized. There would be little likelihood of individual-level data, medical information, and other personal information falling into the hands of people who might attempt to re-identify a data source and use the data and information to the source's detriment. On the other hand, if researchers do not adhere to the policy, it might be providing false security to data sources and policymakers; putting barriers in the way of good research while providing little benefit to data sources.

CONTROLLED ACCESS: PROTECTING WHOM?

Data in dbGaP and other similar repositories are available to essentially any scientist, anywhere in the world, so long as she or he has a graduate degree. I am interested in what happens once controlled-access data leave the repository and reside on data users' computers. The central research question for my pilot study was whether and to what degree controlled-access regimes protect data sources' interests once data have been downloaded to computers at users' facilities. There are good reasons to think that controlled access does not provide the level of protection that policymakers or data sources expect. Primary among these reasons is the fact that NIH and others who operate genome data repositories have no mechanisms for detecting rule violations or enforcing restrictions once data have been downloaded by a user. Considering NIH's GWAS policy, the agency has no mechanism for auditing data users, nor does it require that a user's home institution conduct any audits of research practices or data security. NIH's ability to detect rule violations is haphazard, relying on word of mouth and researchers who self-report. If rule violations are brought to NIH's attention, the GWAS policy contains no enforcement provisions and it fails to specify any penalties or disciplinary procedures. Of course, people comply with rules for many reasons other than fear of penalties, and I am interested in other, extralegal factors that motivate researchers to comply with or violate data use restrictions.

Although the GWAS policy has no formal enforcement mechanism, NIH's status as the major source of funding for academic biomedical researchers in the United States means that an implicit threat to researchers' careers always lurks in the background. Respondents in my study frequently mentioned potential loss of funding as a major reason to comply with any NIH-promulgated rule or requirement, regardless of their views on its necessity or reasonableness. They perceive some real power behind funders' data use rules. As a legal matter, without formal enforcement provisions in the GWAS policy it might be difficult for NIH to justify a decision to withhold funding from a researcher whose grant proposals otherwise met all criteria for funding. Researchers, however, are generally unaware of such legal nuances and their attitudes and motivations are shaped by the threat they perceive.

Implicit threats to funding are not enough to ensure compliance. Such threats are unlikely to have much effect if numerous violations occur and go unpunished. Also, access to dbGaP and the controlled data therein is not limited to academic or publicly funded scientists, nor is

it limited to US-based scientists. Researchers from private firms, large and small, can request data from the repository, as can foreign scientists, who are rarely funded by NIH. The wide availability of controlled-access data serves NIH's aims of leveraging the public's investment in genomics. However, such wide availability means that not all data users will be influenced by the prospect that a violation of data access rules could jeopardize their funding.

Data Use Certificates are a policy innovation and could provide an enforcement mechanism. The certificate is a contract signed by NIH, the requesting scientist (data user), and an official at the scientist's home institution. A party to a contract can sue if the contract's terms are breached, so in theory NIH could enforce rules meant to protect data sources by suing for breach of contract if data users ignored those rules. In practice, NIH is exceedingly unlikely to bring such a suit. Litigation is expensive, and proving the breach might be too difficult. Litigation has fallout, and could derail the careers of productive scientists who were not at fault. Finally, it is not clear how the damages would be calculated—how can the court monetize a scientist's failure to follow an administrative rule? The breach may result in harm or disrespect to data sources, but data sources are not parties to the contract between NIH and the data users. It is not clear what legally cognizable harm the agency would suffer, but the damages would likely be small while the social rupture generated by NIH suing a scientist would be enormous. Furthermore, it is not clear that US courts would enforce the Data Use Certificates, and even less clear whether NIH could get jurisdiction over foreign data users. For all these reasons, Data Use Certificates are more in the nature of symbolic contracts than enforceable ones.

In my study, every data user interviewed could describe one or more instance in which someone else had failed to comply with controlled-access policies. Researchers learned of these breaches by attending colleagues' or students' talks, and by word of mouth. The most common violations involved a researcher with approval to use the data for one particular experiment but using the data in a different, unapproved experiment, and researchers giving unauthorized personnel access to a dataset. For the most part, the unauthorized activities could have been authorized, and some were authorized after the fact (the approving Data Use Committee was unaware the use had already occurred). Every respondent noted that the rule violations involved scientists doing credible research projects and wanting to advance knowledge—not people who were attempting to reconstruct identities and use repository data for nefarious purposes.

Violations of data use rules tend to happen through researchers' inattention, and because the precautions required by these rules run counter to some aspects of laboratory culture and day-to-day practice. The following offers three representative descriptions of rule violations and how they occur:

> [T]here have been people who have, you know, at one point or another inadvertently used data in an analysis that was, you know, for which some of the samples had consent that they could only be used in another kind of analysis, or something like that. But not, I think, out of any ill intent. Just sort of, you know, somebody's post-doc didn't understand that this data wasn't actually available for them to include in their analysis. And those get, you know, internally corrected very quickly … . I can think of a few cases where, you know, somebody's student, you know, didn't check or, you know, two students from different groups were working on something and combined some data without checking.
>
> I think what happens is that you can't help but discuss aspects of analyses you're doing with colleagues. I don't know whether that's considered violating the policy or not … . There's this thing where it's like, "I'm going to take this information out

> of this protected space, and I'm not going to be the only person that's going to touch it." … [V]ery quickly that data needs to be able to be used by other people in order for you to even do your research in your home institution … . I mean, there's always casual exchange between scientists and sometimes people do not wait for somebody's name to get on the data request."
>
> There's, you know, obviously a ton of bleed-through around the edges … .

Of course, researcher intent is not the metric by which one should judge the significance of data use violations. Risk or actual harm to data sources, and failure to respect research participants, are the more important evaluative considerations. To date, there is no evidence that anybody has actually been harmed by an unauthorized use or disclosure of a genome research dataset.

Data users judged known violations of data access rules as harmless, as putting data sources in no real jeopardy. They viewed the violations as *malum prohibitum* (wrong because it is prohibited) but not *malum in se* (wrong in itself, or morally wrong). Respondents were uniformly of the opinion that, for now, it would be exceedingly difficult for a nonexpert to use a research dataset to link sensitive information back to particular individuals. Someone attempting to re-identify data sources would likely need an unusually high level of computational and mathematical skills, and strong motivation to seek the information. Several respondents noted that usually there are easier ways for nonresearchers, including hackers, to obtain the same or equally valuable information about a data source. Some data users assumed that even if it were relatively easy to re-identify data sources, nobody was likely to bother. One respondent stated that an employer or other nonscientist was unlikely to do straightforward Internet searches to help re-identify somebody in a dataset.

Some data users in my sample displayed "doublethink" about identifiability and anonymity similar to the apparent contradiction in the GWAS policy. For instance, one respondent noted that "everyone is completely identifiable from their GWAS signature … . [Y]ou can't be more identifiable than completely identifiable, so, in a sense, there's not more identifiable information in the sequence [than in a genotype]." Yet, when discussing risks to data sources the same scientist stated, "I wouldn't have any concerns about putting my family's data in dbGaP, because … personally I don't perceive any risk in doing that … . [W]hat specific risks to individuals are there … from having their data anonymously in dbGaP?" He describes the type of genome data in repositories as "completely identifiable," and a few minutes later in the interview he describes it as anonymous.

One possible explanation for this apparent contradiction is that genome scientists recognize the genetic data they handle as extremely effective for individuating people, as an identifier, but also feel that society is still in a period when almost nobody can use that identifier. It is as if datasets contained each source's Social Security number, but outside of a very few scientists nobody could connect Social Security numbers to other information about people, such as their names, contact information, or financial records. The potential for re-identification to become much more straightforward and common in the future was a prominent theme in nearly all of the interviews.

Data users were quite "genocentric" when evaluating the risks associated with unauthorized uses of data in genome research repositories —they were contemplating the probabilities that genome information alone could be used to re-identify somebody. They rarely discussed whether the accompanying medical, geographic, behavioral, and other information in a dataset could be used to re-identify somebody. This was true even though all of the interviewees used datasets that contained some medical or other personal information. When pressed on this issue respondents either stated that they had not thought too much about it,

were unsure, or that the data deposition rules of dbGaP and similar repositories would prevent the inclusion of data elements that could be used to re-identify somebody. In theory, this last point is correct. The data deposition rules ought to provide an important layer of protection for data sources. However, nobody has systematically determined whether these rules are being followed and whether they do, in fact, prevent re-identification.

Despite their views that most genome research data in repositories currently carries little risk of harming sources, many data users reported taking significant steps to ensure that their personnel complied with data use restrictions. These steps included developing easily accessible files with information about the use requirements for each dataset in the laboratory; assigning one senior researcher in the laboratory formal responsibility for each dataset, which included ensuring that each person who accessed the data was authorized; and having people outside of the user's laboratory vet every legal and ethical aspect of each dataset that entered the institution (including IRB approvals, consent forms, and Data Use Certificates) to determine whether the dataset could be used and by whom. Some groups had quite sophisticated computer security. Furthermore, several data users spontaneously discussed working to build a culture of appropriate data use within research groups, for instance:

> We've made considerable effort to, you know, continue to, you know, have this [data access and data security] be part of the normal dialog around their projects, and so it now has become over the last few years part of the culture of the institution that everyone's, you know, very carefully attuned to. And the institution, you know, needed to make an investment in being attuned to it … .

Aside from potential threats to their funding, it is worth considering why scientists who think the risks of data sharing are low bother to follow the rules and why they invest in developing a culture of compliance. Data access processes and rules can place significant burdens on scientists, which sometimes dissuade them from seeking datasets. What motivates them to accept these burdens?

One reason for compliance may be that data users view the rules as generally fair, even if they are burdensome and, in many users' opinions, too cautious. The GWAS policy, for example, was developed with consultation and input from the relevant scientific community, and justifications for the rules are well articulated in the *Federal Register* and in other NIH publications. Every user I interviewed was familiar with these justifications and saw some merit to them. Furthermore, the rules have been modified to make inter-institutional collaborations easier. Several respondents pointed to these rule modifications as evidence of flexibility and reasonableness on the part of funders and repositories. Repository rules and processes for data access have, at least for now, a much higher degree of legitimacy among the "regulated parties" than does the Common Rule and its IRB review process.

A second, obvious reason for compliance is that people who obtain data through a controlled-access mechanism get the substantial benefit of access to other people's data. They often get access before the data have been published, because NIH or the Wellcome Trust may require "prepublication data release" as a condition for funding very expensive projects (Toronto International Data Release Working Group 2009). Data users spoke in glowing terms of how mechanisms and institutions that foster data sharing had energized biology research and catalyzed great breakthroughs. They spoke of how their funders' emphasis on data sharing had made scientific collaborations "much more meaningful than they were ten years ago." Of course, I interviewed people who had successfully obtained, used, and published papers based on controlled-access data. This sample was, without doubt, biased in favor of data sharing generally and had experienced a high level of benefit to offset

any bureaucratic burdens. Such researchers are perhaps most likely to comply with data access restrictions so as not to jeopardize future opportunities to obtain data.

A somewhat unexpected reason scientists comply with data access restrictions is that they view the access mechanisms and rules as protecting them from career-ending catastrophes, or preventing them from unintentionally inflicting harm on data sources. Although data users think the risks to sources are currently low in most cases, nearly all scientists I interviewed felt those risks would substantially increase in the future. Furthermore, none of them thought the current risk was zero. Several interviewees expressed concern that a low probability but high-impact unauthorized release of data could have quite negative consequences for the researcher. Following the rules offers a degree of comfort that one is taking reasonable precautions and behaving in an ethically defensible manner. Some researchers, as in the following two cases, find it reassuring that other people—data submitters and repository personnel—are vetting the data, making sure that they do not accidentally contain information that could quite easily identify a source:

> So working through that mechanism [of controlled access], I think that's one of the benefits of it, is at least you've, you kind of feel like, "Okay, I've taken some real steps to be sure that whatever I'm doing, in terms of using this data, I'm doing it in a way that's, you know, that is being intelligent and respectful of, you know, people's privacy and so forth." Because it is a challenging issue … . And it's, like, "I don't want to find out the hard way!" with a knock on the door that like, you know, you've done something really terrible. So, so it is good, I mean, that is one of the advantages of, sort of, working through something like dbGaP, is that you know that you've, at least you've made that attempt to, kind of, work responsibly with the data.

> If there exists an Excel table that will tell me who this is, I shouldn't look at it, or shouldn't get it, or shouldn't share it … . [O]ne thing I find reassuring about dbGaP is at least I don't have to think about [receiving identifying information]. With other data I worry that I might get something that lets me know, and that I would share it without thinking about it … . And so there is, there is a definite reassurance in knowing that, that there is a level of control that's outside of my hands, that's being applied to the data to prevent that.

The fear that bad things will happen to researchers if bad things happen to data sources means that data users' interests align somewhat with those of data sources, despite the lack of any interpersonal interaction or relationship between them.

A related theme is that some data users engage in a complicated but fairly explicit weighing of the burdens and benefits of rule compliance. They have their own ideas about which datasets are riskier than average for data sources, and therefore riskier for researchers to use. Data users' judgments on the risks associated with any particular dataset might not always coincide with those of policymakers, but when researchers view the datasets as risky they are more motivated to abide by data use restrictions, as in these two examples:

> [T]here's an element of self-preservation, like I, I don't want to be the one who's responsible for a security breach, so somewhere in the back of my head is a calculator that's, you know, what is the cost-benefit ratio of the likelihood of something bad happening, versus the inconvenience of going through one or two or five more hoops? And I think that's true for everyone … .

> [F]olks that I've interacted with are generally aware of the degree of privacy of the data that you're dealing with. If it is data that is honest-to-goodness, if it would be easy to get personal information from it, then people do treat it more, more

> tenderly, more responsibly. The further away from that you get the more cavalier folks are about following those specific usage guidelines, obviously. I mean, that's nothing surprising.

In addition to the possibility of source identification, the sensitivity of medical conditions or traits under study, also influenced researchers' motivations to comply with data use restrictions. Types of data respondents described as sensitive included diagnoses of sexually transmitted diseases, information about past surgeries, diagnoses of mental illnesses or information about mental health status, and information that a person is colonized with antibiotic resistant bacteria. One respondent stated that genome research datasets generated from people participating in clinical trials might be so sensitive in the eyes of research participants that they could never be shared through current controlled access mechanism. Most of the information mentioned could be viewed as stigmatizing. Needless to say, researchers' lists of sensitive data were overlapping but not identical.

CONCLUSION

Like Internet advertisers, biomedical researchers want access to large quantities of information about us, including but not limited to our whole-genome genotypes or DNA sequences. Some data in research repositories, including whole-genome data and complex medical information, have the potential to identify people. Whole-genome data are intrinsically identifying, but most researchers believe that the likelihood of anybody using genome information to re-identify somebody in a research dataset is very low, for now. Uncertainty about the ease with which a data source could be identified by genome research data, and about the likelihood of harm, led research funders to create a new type of institution—a controlled-access data repository—to both promote data sharing and protect data sources.

Researchers accept some limits on their ability to share others' personal information and genome data. Researchers' willingness to abide by limits, and to develop norms and practices that implement the limits, depends in large measure on the degree to which they think the data pose risks to data sources. They perceive the current risks as low, and so at least some researchers are careless in complying with data use restrictions. However, in part because compliance helps prevent them from making potentially career-ending data disclosures, researchers find reassurance in using controlled access data from repositories. This reassurance provides some motivation for researchers to comply with burdensome restrictions.

Thus far, researchers' noncompliance has not resulted in any known harm to data sources. My pilot study was not quantitative and could not determine the frequency with which researchers violate data use restrictions. However, if there truly is "a ton of bleed through around the edges," it raises the possibility that our research governance mechanisms ought to be reconsidered.

REFERENCES

Agre P. The Architecture of Identity: Embedding Privacy in Market Institutions. Information, Communication and Society. 1997; 2:1–25.

Berger, E. DNA Sequencing Enters the Terabase Era. Houston Chronicle. http://blog.chron.com/sciguy/2008/07/dna-sequencing-enters-the-terabase-era/

Consortium TGP. A Map of Human Genome Variation From Population-scale Sequencing. Nature. 2010; 467:1061–1073. [PubMed: 20981092]

Contreras J. Bermuda's Legacy: Policy, Patents, and the Design of the Genome Commons. Minnesota Journal of Law, Science and Technology. 2011; 12:61–125.

Department of Health and Human Services (DHHS). Standards for Privacy of Individually Identifiable Health Information. Code of Federal Regulations. 2002; 45 sections 160 and 164.

Department of Health and Human Services (DHHS). Research Repositories, Databases and the HIPAA Privacy Rule. National Institutes of Health; Bethesda: 2004.

Department of Health and Human Services (DHHS). Protection of Human Subjects. Code of Federal Regulations. 2005; 45 section 46.

Department of Health and Human Services (DHHS). Protection of Human Subjects. Code of Federal Regulations. 2005; 45 section 46.101(f).

Homer N, et al. Resolving Individuals Contributing Trace Amounts of DNA to Highly Complex Mixtures Using High-Density SNP Genotyping Microarrays. PLOS Genetics. 2008; 4:e1000167, 1–9. [PubMed: 18769715]

Lowrance WW, Collins FS. Identifiability in Genomic Research. Science. 2007; 317:600. [PubMed: 17673640]

McGuire AL, Gibbs RA. No Longer De-Identified. Science. 2006; 312:370–371. [PubMed: 16627725]

Ohm P. Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization. UCLA Law Review. 2010; 57:1701–1777.

Ostrom E, Hess C. A Framework for Analysing the Microbiological Commons. International Social Science Journal. 2006; 58:335–349.

Ozenberger, B. TCGA: A Future Arrived. National Human Genome Research Institute; http://cancergenome.nih.gov/researchhighlights/leadershipupdate/ozenberger

Strauss, AL. Qualitative Analysis for Social Scientists. Cambridge University Press; Cambridge: 1987.

Toronto International Data Release Working Group. Prepublication Data Sharing. Nature. 2009; 461:168–170. [PubMed: 19741685]

Turnbaugh PJ, et al. The Human Microbiome Project. Nature. 2007; 449:804–809. [PubMed: 17943116]

Zerhouni EA. Policy for Sharing of Data Obtained in NIH Supported or Conducted Genome-Wide Association Studies (GWAS). Federal Register. 2007; 72:49290–49297.

Zweiger, G. Information, Anarchy and Revolution in the Biomedical Sciences: Transducing the Genome. McGraw-Hill; New York: 2001.