Published in final edited form as:

Am J Bioeth. 2008 October ; 8(10): 75–76. doi:10.1080/15265160802478511.

Identifiability of DNA Data: The Need for Consistent Federal Policy

Amy L. McGuire

Baylor College of Medicine

Biological samples are routinely collected and used in biomedical research. As Weir and Olick (2004) point out in their book *The Stored Tissue Issue*, there are four ways in which samples can be stored: “identified” (the name of the person from whom the sample came is included with the tissue); “linked or coded” (the sample has a numerical code that is linked to the name of the person from whom the sample came, most often by computer); “anonymous” (the sample is originally collected without any identifiers linking it to the person from whom it came); and “anonymized” (the sample was originally identified but the identifiers have been irreversibly stripped or disguised) (40). De-identification via linkage, coding, or anonymization has traditionally been considered sufficient to protect privacy, and special protections are only afforded for data or specimens that are readily identifiable. Consequently, research conducted by an investigator who obtains coded or anonymized biological specimens is not considered human subjects research and is not regulated under U. S. federal regulation (the Common Rule) (Federal Policy for the Protection of Human Subjects 2008; Office of Human Research Protections, U.S. Department of Health and Human Services 2004). Similarly, coded or anonymized biological specimens, in and of themselves, are not considered individually identifiable protected health information under the U. S. Health Insurance Portability and Accountability Act of 1996 (2002).

We have challenged this regulatory distinction in the context of genome research, arguing that it is conceptually flawed because DNA is itself uniquely identifiable (McGuire and Gibbs 2006a, 2006b). In 2004, Zhen Lin and colleagues illustrated that access to just 30–80 statistically independent single nucleotide polymorphisms (SNPs) was sufficient to uniquely identify an individual (Lin, Owen, and Altman 2004). Last week, David Craig and colleagues demonstrated that an individual’s SNP profile could potentially be identifiable even when it is aggregated with 1,000 or more other samples (Homer et al. 2008). There have been several policy responses to these studies. Data security has become a top priority for most institutions as well as government agencies. There has also been a shift away from policy that calls for the full public release of all generated sequence data (National Human Genome Research Institute 2003). Newer policies call for the release of data into databases with restricted access and include heightened requirements for informed consent (10). Finally, many individuals and groups have advocated strongly for federal legislation that would protect against the discriminatory use of genetic information in insurance and employment, which recently resulted in the enactment of the Genetic Information Nondiscrimination Act of 2008 (HR 493, Genetic Information Nondiscrimination Act of 2008). Implicit in these policy responses is the

Address correspondence to Amy L. McGuire, Baylor College of Medicine, One Baylor Plaza, Suite 011D, Houston, TX 77030. amcguire@bcm.tmc.edu.

Disclosures: Amy McGuire is a member of the NIH 1000 Genomes Project, Samples/ELSI Workgroup; International Human Microbiome Consortium, Data Release Working Group; NIH Electronic Medical Records in Genomics Network (eMERGE), ELSI Consent and Community Participation Workgroup; NIH Advisory Committee to the Director (ACD), Working Group on Participant and Data Protection (PDP) for the Genetic Association Information Network (GAIN); and the Personalized Health Care Working Group (PHC), American Health Information Community (AHIC), Office of the National Coordinator for Health Information Technology (ONC HIT), Secretary of the Department of Health and Human Services (DHHS). This commentary reflects the views of Dr. McGuire and may not reflect the views of the NIH, DHHS, or other members of these committees.

recognition that DNA data are potentially identifiable and contain sensitive health information and are therefore deserving of special protection. Yet, policymakers continue to promote an interpretation of identifiability that excludes coded and anonymized biological specimens and DNA data from existing regulatory protections (Lowrance and Collins 2007).

In this issue of the *American Journal of Bioethics*, Sara Chandros Hull and colleagues (2008) offer a second reason for challenging this identifiability distinction: it does not seem to be consistent with patient preferences. In a survey of patients from academic medical centers, Hull and colleagues report that 72% of respondents felt that it was important for them to be informed about research using samples that were anonymized (“your name is removed from both the blood sample and from the information from your medical records so you cannot be identified by any of the researchers or anyone else”) and 81% felt it was important to be informed of research using samples that were coded, or “identifiable” (“your name will be replaced with a unique identification number that could be traced back to you and your medical records, if the researcher needs to do so”). Curiosity-based reasons were most commonly identified for those who wanted information about the research use of anonymized samples, while confidentiality concerns were more prevalent among those who only cared to know about the research use of coded samples (Hull et al. 2008). It is not clear whether the potential identifiability of anonymized samples was explained to participants or if they fully understood the privacy risks associated with each scenario. Regardless, as the authors point out, this desire for information about the future research use of biological specimens is consistent with other studies. An important additional contribution of this study lies in the distinction that is made between desire for information and desire for control over decision making. Although a majority of participants expressed a desire for control over decision making (requiring permission to use), 42–43% felt that notification was sufficient, regardless of the identity status of the sample. This suggests an important compromise that deserves further policy consideration. As Hull and colleagues (2008) point out, however, patient preferences are only one factor to be considered in developing responsible research policy.

If conceptual discrepancies and patient preferences are not enough to convince, there is a third reason to abandon the identifiability distinction: it has led to inconsistent federal policy, where DNA data are treated as identifiable for some purposes and not others. For example, the new Policy for Sharing of Data Obtained in NIH Supported or Conducted Genome-Wide Association Studies (GWAS) calls for the release of all genotypic and some phenotypic data from NIH funded or supported GWAS into dbGAP, a restricted database maintained by the NIH (National Institutes of Health 2007). Depositing investigators who obtain the data either through direct interaction with participants or who obtain identifiable private information (linked to the subject’s name or other personally identifying information) are assumed to be conducting research involving human subjects and must obtain informed consent and IRB approval for the study and the subsequent data release. However, investigators who desire access to the data in dbGAP, which are coded or anonymized by the depositing investigator, are not conducting human subjects research and do not need IRB approval for their study. They need only to obtain approval from the NIH data access committee (DAC).

At the same time, there is concern that since the NIH is a government agency data stored in dbGAP could be accessible to the public subject to a request made under the Freedom of Information Act (FOIA). To protect against this threat, the NIH maintains that because DNA is uniquely identifiable, complying with a FOIA request for that information would constitute a “clearly unwarranted invasion of personal privacy” and thus, the data in those databases are exempt from FOIA (Lowrance and Collins 2007). Regardless of the merits of this claim, it is problematic to treat DNA data as identifiable for the purposes of FOIA but not identifiable under the Common Rule. Conflicting interpretations of the identifiability distinction leads to incompatible policies that are confusing to investigators, IRBs, research participants, and even

polymakers themselves. Worse, inconsistent policies are more susceptible to challenge. For example, the NIH's assertion that the data in dbGAP are exempt from a FOIA request could be challenged in a court of law. There is no judicial precedent on this issue. Will the courts believe that coded DNA data are identifiable (making access under FOIA a "clearly unwarranted invasion of personal privacy") when the Office of Human Research Protections (OHRP) has issued federal guidance to the contrary? Inconsistent interpretations of the identifiability of DNA data at a federal policy level threatens to discredit legitimate attempts to protect these data, which may ultimately undermine public trust and inadvertently harm research participants.

Acknowledgments

This publication was made possible by the Greenwall Foundation Faculty Scholars Program in Bioethics and NIH/NHGRI-ELSI R01 HG004333.

References

- Federal Policy for the Protection of Human Subjects. 2008. Code of Federal Regulations Title 45, section 46.
- Health Insurance Portability and Accountability Act of 1996. 2002. Security and privacy. Code of Federal Regulations Title 45, section 164.
- Homer N, Szelinger S, Redman M, Duggan D, Tembe W, Meuhling J, 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.
- Hull SC, Sharp RR, Botkin JR, et al. Patients' views on identifiability of samples and informed consent for genetic research. *American Journal of Bioethics* 2008;8(10):x–xx.
- Lowrance WW, Collins FS. Identifiability in genomic research. *Science* 2007;317:600–602. [PubMed: 17673640]
- Lin Z, Owen AB, Altman RB. Genomic research and human subjects privacy. *Science* 2004;305:183. [PubMed: 15247459]
- McGuire AL, Gibbs RA. No longer de-identified. *Science* 2006a;312:370–371. [PubMed: 16627725]
- McGuire AL, Gibbs RA. Meeting the growing demands of genetic research. *Journal of Law, Medicine, and Ethics* 2006b;34:809–812.
- National Human Genome Research Institute. Bethesda, MD: National Institutes of Health; 2003. Reaffirmation and Extension of NHGRI Rapid Data Release Policies: Large-Scale Sequencing and Other Community Resource Projects. Available at www.genome.gov/10506537
- National Institutes of Health. Bethesda, MD: National Institutes of Health; 2007. Policy for Sharing of Data Obtained in NIH Supported or Conducted Genome-Wide Association Studies. Available at <http://grants.nih.gov/grants/guide/notice-files/NOT-OD-07-088.html> HR 493, Genetic Information Nondiscrimination Act of 2008. Available at <http://www.govtrack.us/congress/billtext.xpd?bill=h110-493>
- Office of Human Research Protections, U. S. Department of Health and Human Services (HHS). Washington, DC: HHS; 2004. Guidance on Research Involving Coded Private Information or Biological Specimens. Available at www.hhs.gov/ohrp.humansubjects/guidance/cdebiol.pdf
- Weir, RF.; Olick, RS. *The Stored Tissue Issue: Biomedical Research, Ethics, and Law in the Era of Genomic Medicine*. New York, NY: Oxford University Press; 2004.