ETHICS

Identifiability in Genomic Research

Genomic data are unique to the individual and must be managed with care to maintain public trust.

William W. Lowrance and Francis S. Collins

Genomic research can now readily generate data that cover significant portions of the human genome at levels of detail unique to individuals. Data can now be categorized with respect to disease-related genes and linked to clinical, family, and social data. Identifiability, the potential for such data to be associated with specific individuals, is therefore a pivotal concern. Research, health care, police, military, and other DNA and genotype reference collections are growing. Members of the public and its leaders worry about risks of erroneous or malicious identity disclosure and consequent embarrassment; legal or financial ramifications; stigmatization; and/or discrimination for insurance, employment, promotion, or loans.

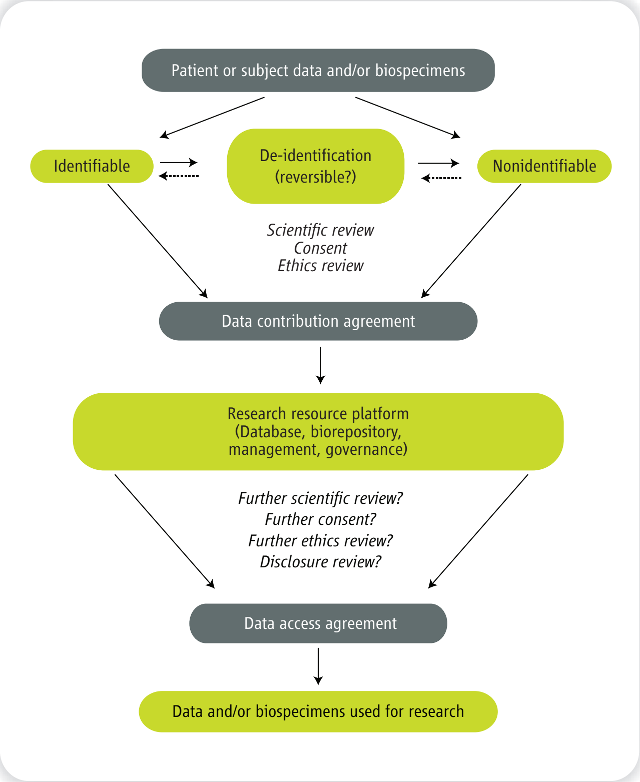
If the data are considered identifiable, they may be covered by informational or genetic privacy laws, with implications for consent and other rights. They may be covered by human-subjects regulations, with implications for oversight. Controlled, conditional release may be required for the data as opposed to open public release. These can all be obstacles to the conduct of health-related research.

In the United States, personal data used in health care and/or research are protected by the Common Rule on Protection of Human Subjects (*1*), and the Privacy Rule under the Health Information Portability and Accountability Act (HIPAA) (*2*–*4*). They are also protected by state and other federal laws and regulations. In the European Union (*5*), informational privacy is protected by national laws that implement the Data Protection Directive, such as the U.K. Data Protection Act (1998). Most other countries have similar laws.

How these laws apply specifically, and how adequate they are in the genomic research arena, is not entirely clear. Protection of privacy was among the issues examined by the National Institutes of Health (NIH) in a recent public consultation (*6*).

New Modes of Data Flow

Until recently, most genomic research used data and biospecimens obtained fairly directly, from the data subjects themselves or clinical repositories or specialized research collections. This will continue, as it has many advantages. But now, in efforts to increase the range and quantity of data, large-scale research platforms are being built that assemble, organize, and store data, and sometimes biospecimens, and then distribute these to researchers (see figure). The advantages of such platforms, in addition to scale, are that they can be a robust staging-point for screening data quality, fostering uniformity of data format, and facilitating analysis. Some platforms accumulate data directly (as the Framingham Heart Study does); others assemble them from a variety of sources (as The Cancer Genome Atlas, the Genetic Association Information Network, and the Wellcome Trust Case Control Consortium do and U.K. Biobank will) (*7*). Among the design and governance issues are whether and how to de-identify the data and at what stages to conduct scientific and ethics review.


Patient or subject data and/or biospecimens
Identifiable
De-identification (reversible?)
Nonidentifiable
Scientific review
Consent
Ethics review
Data contribution agreement
Research resource platform (Database, biorepository, management, governance)
Further scientific review?
Further consent?
Further ethics review?
Disclosure review?
Data access agreement
Data and/or biospecimens used for research

These new data flows, genomewide analyses, and novel arrangements such as the Informed Cohort scheme recently proposed by Kohane *et al*. (*8*) are relatively uncharted territory with respect to human subjects and privacy considerations. Precedent doesn't provide sufficient guidance. For example, the Human Genome and HapMap Projects have genotyped DNA from only a few hundred carefully selected people who prospectively consented to the analysis and to open publication after thorough explanation, discussion, and community consultation. The projects have been scrutinized closely all along. But when the data relate to more people (by orders of magnitude) or to retrospective analysis of biospecimens, then for pragmatic reasons such painstaking selection, consent negotiation, and scrutiny can't generally be achieved.

Identifiability and Identifiers

Identifiability ranges from overtly identifiable, to potentially identifiable by deduction, to absolutely unidentifiable. The concept isn't simple, as evidenced by the European Commission's publication of an elaborate "Opinion on the concept of personal data" in June 2007, 12 years after passage of the Data Protection Directive (*9*).

In legal regimens, indirect identifiability is as important as direct. For instance, the HIPAA Privacy Rule applies to "information that identifies an individual; or with respect to which there is a reasonable basis to believe the information can be used to identify the individual" (Sec. 160.103). Similarly, the U.K. Data Protection Act applies to "data which relate to a living individual who can be identified—(a) from those data, or (b) from those data and other information which is in the possession of, or is likely to come into the possession of, the data controller" [Sec. I.1-(1)]. If

W. W. Lowrance is a consultant in health research policy and ethics, 72 rue de St. Jean, CH-1201 Geneva, Switzerland; e-mail: lowrance@iprolink.ch. F. S. Collins is director, U.S. National Human Genome Research Institute, Bethesda, MD 20892–2152 USA; e-mail: francisc@mail.nih.gov

Downloaded from www.sciencemag.org on August 24, 2007

Published by AAAS

data aren't identifiable they shouldn't be considered "personal," and a variety of rights and obligations that apply to personal data may not be relevant.

Three sorts of identifying factors can be distinguished: demographic or administrative tags (e.g., name, social security number, e-mail address, hospital name, postal code); overt descriptors (e.g., gender, eye color, height, blood type, scars, asthma); and indirect clues (e.g., medication use, number of children, spouse's occupation, circumstances of emergency-room admission). Whether particular bits of data alone or in combination should be considered sufficient to identify a person is a matter of judgment. Much may depend on whether partial identifiers can be linked with identified or identifiable data in public or other databases.

The HIPAA Privacy Rule illustrates the practical challenges. For data to be considered adequately de-identified and therefore not subject to its provisions, a number of descriptors, which it lists, must be absent [Sec. 164.514(B)(2)] (*7*). The list contains identifiers that are linked fairly directly to name and address, such as medical record numbers or hospital discharge dates. Knowing a few elements on the list may or may not allow identification, and even knowing a person-unique fact such as social security number allows identification only if it can be traced to the person through some other source.

Identifying Through Genomic Data

Matching against reference genotype. The number of DNA markers such as single-nucleotide polymorphisms (SNPs) that are needed to uniquely identify a single person is small; Lin *et al.* estimate that only 30 to 80 SNPs could be sufficient (*10*). Thus, such data can be used, with high certitude, to confirm that two samples come from the same person; whether this can identify anybody in the usual sense depends on whether the reference data are personally identified.

Collections that can be used for matching continue to grow. Identified biospecimens from millions of people are held by criminal justice systems and armed services (*11,12*). Biospecimens and a growing number of genomic analyses are held by health-care, public health, and health research institutions. To be clear, the risk is not that a match might be found but that a de-identified data set will become linkable to a specific person because the matched data set contains personal identifiers.

Linking to nongenetic databases. A second route to identifying genotyped subjects is deduction by linking and then matching genotype-plus-associated data (such as gender, age, or disease being studied) with data in health-care, administrative, criminal, disaster response, or other databases (*10,13,14*). There is no shortage of public and commercial databases about people's lives, especially in the United States. If the nongenetic data are overtly identified, the task is straightforward . Even if such data are not fully identified, inferential narrowing-down may be possible. Statisticians have many techniques for identifying data subjects from partial data (*15,16*).

Profiling from genomic data. A number of physical attributes can now be inferred from DNA analysis, such as gender, blood type, approximate skin pigmentation, and manifestations of Mendelian disorders. Reliability of predictions will likely increase regarding height or other aspects of skeletal build, hair color and texture, eye color, and even some craniofacial features. Soon many chronic disease susceptibilities will be predictable and, before long, some behavioral tendencies will be. In 5 to 10 years, many attributes will be profilable.

Tactics for De-identifying Genomic Data

Limiting the proportion of genome released. The first option is to release only limited segments of genomes, such as sequence traces or a few variants, along with minimum necessary phenotypic or other data. But "how much" is sufficient for identifying, by any route, depends on the region and extent of genome covered, the density of mapping, the rarity of variants, the degree of linkage disequilibrium, and other factors (*17*). This makes it difficult to develop general guidance on how much to expose publicly.

Many projects do limit the portion of genome they release, especially if the release is unrestricted. Precautions can be taken, such as releasing sequence traces in such a separated manner that no individual's data can be reassembled by overlaps. But releasing too-few SNPs or too-short snippets of sequence may thwart research.

Statistically degrading data. This is possible, for example, by lumping all purines and all pyrimidines. Unfortunately, the occurrence of a T instead of a C in one data cell can mean the difference between disease and health. So for many lines of genomic research, degrading data degrades usefulness.

Sequestering identifiers via key-coding (reversibly de-identifying) (*7*). This is the method most widely used in health research. Administrative or other overt identifiers are separated from data, but a link is maintained between them via an arbitrary numerical key-code (*18*). Held securely and separately, the key allows reassociation of substantive data with identifiers if necessary. The key and responsibility for its use can be delegated to a trusted party; its use can be guided by agreed-upon criteria and subjected to oversight.

Provision of Access to Data

Open versus controlled release. A cultural habit of rapid, open release of genomic data has been pursued by the involved scientists and institutions since the beginning of the Human Genome Project (*19–20*). There is no question about the research advantages of such principles and policies. But almost certainly, the principles will have to be modified now for databases that include extensive genotypic information, to heighten the protection of identifiability (*21*).

Open data release, as with deposition in a publicly accessible Web site, is acceptable only if either: (i) the data are for all practical purposes not identifiable; or (ii) consent to the release is ethically legitimate and is granted by the data subjects, or the necessity for consent is waived by a competent ethics body. Most projects now take three precautionary steps: sequestering the standard identifiers via key-coding; performing disclosure risk-reduction (such as by rounding birth date to year of birth); and providing access to the de-identified data under conditional terms.

Terms of agreements. Data-access agreements (alternatively called "certifications" or "use agreements") cover many matters. Legally they amount to contracts, and they may have to be entered into by researchers' institutions as well as the researchers.

Agreements may set limitations on purposes and uses, allowed users, or other matters covered by consent, either for the whole dataset or for particular data-subjects, and may address how data will be released. They should refer to physical, organizational, and information technology security. They may specify who will be responsible for de-identifying data and may cover key-coding, safeguarding of the key, and criteria for use of the key. They should always state that researchers will make no attempt to identify nonidentified data. They should restrict unauthorized passing on of data and should extend the chain of custody and the accompanying obligations if data are passed on. They may address linking, if linking to other datasets is contemplated that might increase identifiability. Invariably they require that derived data on individuals be protected at least as carefully as the data being provided. They may make access contingent on Institutional Review Board or other ethics committee approval and may specify the stage(s) at

Downloaded from www.sciencemag.org on August 24, 2007

Published by AAAS

which ethics review should be conducted.

Oversight. Most data-release decisions, including those made by curating principal investigators, are overseen or made by stewardship committees. This not only protects the data subjects, but it tends to maximize data sharing and to protect investigators, hosting institutions, research platforms, and funders from perceptions or acts of favoritism or impropriety.

Extremely restricted access. Examples are data enclaves in which certified researchers perform studies in databases on a special server. Because this can prevent users from taking away or sharing data examined or detailed records of the analysis, and can deter scrutiny by coauthors, manuscript reviewers, or medical products regulators, the approach must be used only as a last resort.

Scaling to Risks

Risks to data subjects, to data stewards, to researchers and their institutions, and even to the genomic research enterprise must be examined. The ease of identifying people from DNA or genomic data, without breaking laws, should not be overstated; it takes competence, perhaps a laboratory equipped for the purpose, computational power, perhaps linking to other data, and determined effort. But some risks are real. Data cordoned off and curated for research can be exposed to external view by deliberate transfer; accidental or careless release; theft; release under court order or law-enforcement demand; and release in response to freedom-of-information (FOI) request.

Data must be de-identified proportionate to reasonably expectable risks. The conditions on release should not be so burdensome as to retard research, but they must be binding. Court orders must be honored, but indiscriminate trawling through databases should be discouraged, and compelled genotype releases should be limited to the data actually needed for the investigation.

Construal of genomic "human subject." If data have been de-identified but include large amounts of genetic information, are the individuals still considered "human subjects"? The answer has important implications for consent, ethics review, and safeguards. McGuire and Gibbs have urged that "genomic sequencing studies should be recognized as human-subjects research and brought unambiguously under the protection of existing federal legislation" (*22*), but this could be unnecessarily extreme. In the United States, the Office of Human Research Protections considers that data or biospecimens collected for one purpose but then key-coded and used secondarily for research are not "individually identifiable," and therefore the research is not human-subjects research (*7*). This is a strong incentive to support de-identification and to de-identify data.

Certificates of confidentiality. These are legal assurances that the NIH and some other agencies can issue that "allow the investigator and others who have access to research records to refuse to disclose identifying information on research participants in any civil, criminal, administrative, legislative, or other proceeding, whether at the federal, state, or local level" (*23*). Their use deserves rigorous evaluation, and they may deserve administrative or legislative buttressing.

Sanctions against breach of access commitments. Generally the experience with controlled access has been positive. But the robustness and enforceability of access arrangements will be tested by the increasing provision of data to recipients who have not had prior relationships with the principal investigators who collected the data, the funding agencies, or the centers that distribute the data. Funders can consider rescinding grant support or denying future support, but they have less recourse against breaches by nongrantees. New legal penalties may be needed.

FOI requests. In a number of countries, most information held by government bodies must be made available to the public upon formal request. But there are limits, including protection against invasion of personal privacy. Given that genotype data, even though key-coded and de-identified, might be identifiable under some current or future circumstances, responses to FOI requests should negotiate to release only data relevant to the particular inquiry and to redact the data on individuals to reduce the risks.

Genetic antidiscrimination laws. As a complement to the protections discussed in this article, several countries have adopted or are considering adopting genetic antidiscrimination laws. An example is the Genetic Information Nondiscrimination Act currently under consideration in the U.S. Congress, which prohibits discrimination on the basis of genetic information with respect to health insurance and employment (*24*).

Conclusion

A proper balance between encouraging genomic research and protecting privacy and confidentiality of research participants will not be easily achieved. Only rarely will a completely open access model be defensible when sufficient amounts of genomic data are present to be unique to the individual. A variety of controlled-access models can be utilized, however, that minimally impede access by qualified investigators and at the same time keep the risk of identifying individuals low. Protection of identifiability is obligatory for maintaining the trust of our most important research partners, the public.

References and Notes

1. "Federal policy for the protection of human subjects," 45 Code of Federal Regulations (CFR) §46 (2005); www.hhs.gov/ohrp/humansubjects/guidance/45cfr46.htm.
2. Department of Health and Human Services, "Medical privacy: National standards to protect the privacy of personal health information"; www.hhs.gov/ocr/hipaa.
3. National Institutes of Health, "Protecting personal health information in research: Understanding the HIPAA Privacy Rule"; http://privacyruleandresearch.nih.gov/ pr_02.asp.
4. Congressional Research Service, "Federal protection for human research subjects: An analysis of the Common Rule and its interactions with FDA regulations and the HIPAA privacy rule" (2005); www.fas.org/sgp/crs/misc/RL32909.pdf.
5. European Commission, *Data Protection in the European Union* (EU), www.ec.europa.eu/justice_home/fsj/privacy.
6. Genome-Wide Association Studies, http://grants.nih.gov/grants/gwas/index.htm.
7. Further information can be found in supporting material on *Science* Online.
8. I. S. Kohane *et al.*, *Science* **316**, 836 (2007).
9. EU Data Protection Working Party, "Re: Article 29, Opinions 4/2007 on the concept of personal data," adopted 20 June 2007; http://ec.europa.eu/justice_home/fsj/privacy/docs/wpdocs/ 2007/wp136_en.pdf.
10. Z. Lin, A. B. Owen, R. B. Altman, *Science* **305**, 183 (2004).
11. J. M. Butler, *J. Forensic Sci.* **51**, 253 (2006).
12. J. M. Butler, *Forensic DNA Typing* (Elsevier, Amsterdam, ed. 2, 2005).
13. B. Malin, L. Sweeney, *Proc. J. Am. Med. Inform. Assoc.* **2000**, 537 (2000); http://privacy.cs.cmu.edu/dataprivacy/projects/genetic/dna1.html.
14. B. Malin, *J. Am. Med. Inform. Assoc.* **12**, 28 (2005).
15. Federal Committee on Statistical Methodology, "Report on statistical disclosure limitation methodology"; www.fcsm.gov/working-papers/spwp22.html.
16. American Statistical Association Web site on Privacy, Confidentiality, and Data Security; www.amstat.org/comm/CmtePC.
17. Z. Lin, R. B. Altman, A. B. Owen, *Science* **313**, 441 (2006).
18. W. W. Lowrance, *Learning from Experience: Privacy and the Secondary Use of Data in Health Research* (The Nuffield Trust, London, 2002), especially pp. 32 and 33; www.nuffieldtrust.org.uk/ecomm/files/161202learning.pdf.
19. National Human Genome Research Institute, "Reaffirmation and extension of NHGRI rapid data release policies," 2003; www.genome.gov/10506537.
20. Wellcome Trust, "Policy on data management and sharing "(Wellcome Trust, London, January 2007); www.wellcome.ac.uk/doc_WTX035043.html.
21. W. W. Lowrance, *Access to Collections of Data and Materials for Health Research: A Report to the Medical Research Council and The Wellcome Trust* (Wellcome Trust, London, 2006); www.wellcome.ac.uk/doc_WTX030843.html.
22. A. L. McGuire, R. S. Gibbs, *Science* **312**, 370 (2006).
23. National Institutes of Health, Certificates of Confidentiality Kiosk, 28 February 2006; http://grants.nih.gov/ grants/policy/coc.
24. H.R. 493, S. 358, 110th Congress, 1st Sess. (2007).

25. This article reflects the deliberations of a 3 to 4 October 2006 workshop on identifiability, which built upon a white paper prepared by W.W.L. Both were supported by the U.S. National Human Genome Research Institute. The authors thank the workshop participants for constructive contributions.

10.1126/science.1147699

Downloaded from www.sciencemag.org on August 24, 2007

Published by AAAS