SOUNDING BOARD

# Health-Information Altruists — A Potentially Critical Resource

Isaac S. Kohane, M.D., Ph.D., and Russ B. Altman, M.D., Ph.D.

One of the key ideas behind sequencing the human genome was the promise of "personalized medicine." The idea was that genetic information could be used to make health care more precise, efficacious, and safe. The Human Genome Project showed us that among humans, DNA sequences are 99.9 percent similar, but the remaining 0.1 percent, in the context of environmental and epigenetic factors, produces the entirety of genetic variability within the human population. How can we use the information about human genetic variation to achieve these stated goals of the genome-sequencing effort?[1] Investigators are currently collecting phenotypic information about patients (their disease diagnoses, prognoses, and treatments) and comparing it with their DNA sequences. Methods used to obtain these large phenotypically annotated populations may not be adequately productive because of concerns about privacy and disclosure of genotypic and phenotypic data. We think these concerns are real but addressable sociologically, technologically, and legislatively. The basic idea is that giving patients control of the use of their health data will provide a practical mechanism for harnessing the volunteerism of our populations and gathering research data.

## THE EFFECTS ON CLINICAL GENOMIC RESEARCH OF LARGE POPULATIONS THAT HAVE BEEN PHENOTYPED

Although types of genomic technology such as sequencing and genotyping have become efficient enough to give them commodity status (i.e., cents per genotype or less), the collection of phenotypic data continues to have low throughput and is the most challenging and costly aspect of large studies. Given the limited pool of study subjects, it is difficult (within the current clinical research infrastructure) to execute studies with sufficient statistical power to identify the genotypic basis leading to small phenotypic effects. We need larger study cohorts. For example, Francis Collins, director of the National Human Genome Research Institute, has called for large cohorts (at least 200,000 subjects) to be assembled simply to achieve the necessary sample sizes to overcome the problems of cross-sectional studies.[2]

## CAUSE FOR CONCERN: NO PERFECT ANONYMITY

When they agree to participate in research studies, patients assume that their health history will be used to advance research but will be kept confidential and never used to discriminate against them. As a consequence, researchers invest substantial efforts in removing any information from research data sets that could be used to identify the specific participants.

However, a recent study by Malin and Sweeney concerning database security has shown that apparently de-identified subjects often can be either unambiguously re-identified or partially identified by means of filtering the data to a very small subgroup of potential matches.[3] Malin and Sweeney took publicly available and de-identified hospital-discharge data from Illinois (from 1990 through 1997) and combined them with Census data and voter-registration data to identify patients with rare genetic diseases. They showed that 33 percent of patients with cystic fibrosis could be re-identified, as could 50 percent of patients with Huntington's disease, 70 percent of patients with Fanconi's anemia, and 100 percent of patients with Refsum's disease (a very rare disorder).

The key insights from the work of Malin and Sweeney are that hospital-discharge data are sufficient to associate a given disease with particular features of the DNA sequence, and that discharge data can be combined with other public data to associate the pattern of hospital visits with a patient's home ZIP Code, age, and sex. In these analyses, the combined information becomes a unique or a near identifier of a person and creates the link among DNA sequence, disease, and personal identification,

ultimately raising concern that guaranteeing complete confidentiality may never be possible.[4]

Although Malin and Sweeney focused on rare diseases, the availability of increasing amounts of health information makes everyone rare in some way. Indeed, as reported in an earlier demonstration,[5] Sweeney obtained the health records of former governor of Massachusetts William F. Weld with the use of the most common of data — hospital information about state employees, who were identified only by ZIP Code, sex, and date of birth, published by the Commonwealth of Massachusetts Group Insurance Commission. With the use of a voter-registration list purchased for $20, Sweeney identified three persons with the same date of birth and sex as the governor — and only one who also had the same ZIP Code. The governor's Group Insurance Commission record thereby was identified fully. Even without informative demographic data, the augmented richness of phenotypic characterizations increases the possibility of identifying people. In reaction to the increased availability of information, improved requirements for guaranteed confidentiality could threaten the data-hungry postgenomic scientific agenda.

## THE RISKS OF SHARING DATA

Sharing genetic and health information for research, policymaking, and marketing purposes is undoubtedly associated with some risk. First, the long-term implications of publicly available genetic-sequence data are not fully understood. Data that are released about DNA may be innocuous now but may contain implicit information that becomes apparent with future discoveries. Second, nefarious parties may use data to try to re-identify persons for the purposes of economic or social gain or pure mischievousness. Third, the decision to share genetic data affects immediate family members, extended family, ethnic groups (which may not be comfortable with sharing data), and many other associated groups that share DNA. These associations make individual genetic information communal in a very real sense.

## VARYING LEVELS OF CONCERN

Not surprisingly, the level of concern about these aspects of health privacy varies broadly within the population.[1] Some potential study subjects refuse to have their information included in any database because of privacy concerns, whereas others show only moderate or no concern. The reasons for this variation are not known. Those who show little concern may focus on the prospect of their data contributing to improving human health. They may have reasons (e.g., personal illness or the illness of a loved one) for participating in disease-oriented research and may not be worried about threats to the privacy of their health history. These attitudes may not be illogical, because despite the potential for abuse — and the publicity surrounding some particularly egregious abuses — the record of actual misuse of health information is limited.[6] The existence of potential study subjects who are neutral or enthusiastic about the public sharing of their health data creates a cadre of "genetic-information altruists." They may not constitute a majority of the population, but they exist and represent a valuable resource for researchers of genetics. Whether such volunteers present biases in terms of social milieu, educational background, ethnic background, or race relative to those who have declined to volunteer restates the problem of ascertainment bias among volunteers and study subjects that was recognized as early as 1934 by R.A. Fisher and subsequently addressed in an extensive literature about accounting and compensating for such biases.

There is ample evidence of a range of sensitivities about genetic information. J. Craig Venter disclosed that his DNA was used as part of the private genome-sequencing effort.[7] The group Californians for Universal Voluntary Individual Genome Sequencing has called for genome sequencing of every citizen of California.[8] The group recognizes the need for "proper privacy controls" but sees the benefits in terms of health as a critical impetus.

The governments of Iceland (in partnership with deCODE Genetics) and Estonia have set up national data banks linking genetic and clinical data, with participation by large percentages of their populations. These efforts, in milieus that may have different cultural and economic perspectives, indicate a societal comfort with sharing data.[9] Admittedly, in Iceland there was substantial debate and disagreement about the appropriateness of the detailed rules for inclusion in the national database. Even so, the government modified the plan and moved forward with it. In Estonia, the effort, still in its early development, has benefited from the experience in Iceland, and the national database was introduced with little controversy.

Despite concern about privacy, patients are of-

ten willing to share their genetic data with scientific researchers, even more than with their own relatives.[10] Important exceptions to this willingness exist for patients with mental illness, sexually transmitted diseases, or other diseases that are particularly sensitive or carry a stigma.

Consumers of health care may, in some cases, be ahead of health care providers in terms of knowledge and awareness of genomic information. Many clinicians have insufficient training or understanding of genetic testing to recommend its use or to interpret the results,[11,12] whereas disease-based or gene-based interest groups routinely disseminate relevant genetic information to their constituents.[13]

## A POTENTIAL SOLUTION AND A PROPOSAL

The emergence of patient-controlled electronic medical records, as part of the plan of the Department of Health and Human Services for health-related information technology, provides several potential mechanisms for the disclosure of health information. The patient-controlled medical record exists in parallel to institutional health records and constitutes an electronic copy of all relevant records, with control of disclosure left to the patient.[14,15] Population-wide queries may be made across databases of these patient-controlled records contingent on the patients' having authorized investigators to access all or a portion of their records. The patient-controlled medical record does not require a single central database, thus reducing concern about centralized threats. Most important, personally controlled health records permit the accretion over time of enriched phenotypic data about patients, as the patients interact with different health care providers, without depending on the assent of entities other than the patient. With the emergence of personal health records, patients are able to release precisely the data they are willing to share, and thus the system can accommodate various degrees of information altruism.

These observations lead us to consider the idea that large-scale studies of genotype and phenotype should specifically seek out volunteers who are information altruists. The guarantees made to these subjects about the risks of re-identification can then be more realistic. The potential damages can be outlined, but the subjects presumably will elect to take the risk in the hope of helping to address human disease.

A particular concern is that the use of a subgroup of people who share information might influence statistical analyses of associations. This concern must be addressed with the use of statistical methods for the estimation of bias, as well as by careful study design, perhaps including the use of genomic markers to ensure that genotypic and phenotypic diversity is maintained, just as for any other cohort study. However, the virtue of encouraging maximal information altruism is that more information about each patient makes possible better determination and correction of bias. Databases systematically stripped of data in the often vain hope of preventing re-identification may obscure these biases.

A moderate set of actions by policymakers and legislators could clear the field for these studies. First, rules could be implemented to make it illegal to link health information contained in research databases to other data resources, so as to prevent the inference of individual information outside the scope of the original informed consent. Such a prohibition would complement technological protections to provide protection against damage that could be traced to the use of research databases. The mechanisms of enforcement and the effectiveness of this deterrent can be debated, but a prohibition would clearly make large corporate and governmental entities think twice before aggregating health-related data.

Second, researchers who curate genetic databases should have some protection for their activities, provided that they follow an agreed-on set of operating guidelines. Currently, standards for providing de-identified data are not compatible with recent demonstrations of re-identification with the use of data-aggregation techniques. At the same time, taking all sources of genetic data off-line could severely curtail progress in genetic research. Thus, curators of genetic databases need protection in the event that an otherwise bona fide user abuses the information. A set of standard expectations would not guarantee privacy (a standard that is too high) but instead would mandate reasonable physical and logistical security.

Third, and most important, patients should be granted explicit control over the disclosure process. They should be able to indicate the types of users who can see their data, and they should be able to request lists of those who have seen it. This capability would effectively grant to individuals the task of enforcing the first two policy recommendations.

The next step should be to create pilot studies to test the feasibility of asking patients to accept lower levels of privacy guarantees. These studies would require the development of new consents. They would also limit the data that are disclosed to include only those subsets of information that the patient is willing to share. At the same time, policymakers need to provide some protection for the patients and the researchers. The combination of privacy guarantees that are "good enough" and societal protection of the rights of patients and researchers could hasten the improvements in health care that will result during the post-genomic era.

From the Department of Pediatrics, Children's Hospital and Harvard Medical School, Boston (I.S.K.); and the Department of Genetics, Stanford University, Stanford, Calif. (R.B.A.).

**1.** The International HapMap Consortium. The International HapMap Project. Nature 2003;426:789-96.
**2.** Collins FS. The case for a US prospective cohort study of genes and environment. Nature 2004;429:475-7.
**3.** Malin B, Sweeney L. How (not) to protect genomic data privacy in a distributed network: using trail re-identification to evaluate and design anonymity protection systems. J Biomed Inform 2004;37:179-92.
**4.** Lin Z, Owen AB, Altman RB. Genetics: genomic research and human subject privacy. Science 2004;305:183.
**5.** Sherman E. It doesn't take much to make you stand out. Cambridge, Mass.: Harvard University Extension School Bulletin, Fall 2001.
**6.** Clayton PD, Boebert WE, Defriese GH, et al. For the record: protecting electronic health information. Washington, D.C.: National Academies Press, 1997.
**7.** Wade N. Scientist reveals secret of genome: it's him. New York Times. April 27, 2002:A1, A15.
**8.** Strassman M. New bioactivist group calls for universal, voluntary individual genome sequencing for all California residents. California Politics Today. September 6, 2004.
**9.** Kaiser J. Biobanks. Population databases boom, from Iceland to the U.S. Science 2002;298:1158-61.
**10.** Henneman L, Timmermans DR, van der Wal G. Public experiences, knowledge and expectations about medical genetics and the use of genetic information. Community Genet 2004;7:33-43.
**11.** Sifri R, Myers R, Hyslop T, et al. Use of cancer susceptibility testing among primary care physicians. Clin Genet 2003;64:355-60.
**12.** Genetic testing for breast and ovarian cancer susceptibility: evaluating direct-to-consumer marketing — Atlanta, Denver, Raleigh-Durham, and Seattle, 2003. MMWR Morb Mortal Wkly Rep 2004;53:603-6.
**13.** Barlow-Stewart KK, Gaff CL. Working in partnership with support services in the era of the "new genetics." Med J Aust 2003;178:515-9.
**14.** Simons WW, Mandl KD, Kohane IS. The PING personally controlled electronic medical record system: technical architecture. J Am Med Inform Assoc 2005;12:47-54.
**15.** Mandl KD, Szolovits P, Kohane IS. Public standards and patients' control: how to keep electronic medical records accessible but private. BMJ 2001;322:283-7.