Discussion document for workshop convened by the National Human Genome Research Institute, Bethesda, October 3–4, 2006.

Privacy, Confidentiality, and Identifiability in Genomic Research

William W. Lowrance, Ph.D.

Consultant in health research ethics and policy,
Geneva (lowrance@iprolink.ch)

October 3, 2006

(This document was prepared by the project leader to catalyze and inform discussion in the workshop. It should not be taken to express the views of anyone other than the author.)

Contents

1. *The challenges*

Protect privacy *and* foster research

For reasons that need not be rehearsed here, personal information collected in the course of health care, payment, or research must be protected, and the personal rights and dignity of data-subjects and biospecimen sources must be respected. This obligation prevails in most countries and is enshrined in professional ethics and in law.

(A note regarding scope. This project is oriented to the U.S. situation, but similar issues should be of concern everywhere genomic research is pursued.)

In the U.S., rights relating to personal information handled in health care and/or research are governed by two omnibus regulations, the Common Rule on Protection of Human Subjects (hereafter, "Common Rule")[1] and the Privacy Rule under the Health Information Portability and Accountability Act ("HIPAA Privacy Rule").[2, 3] They are also addressed by many State statutes and regulations, some of which focus specifically on genetic data. Personal data held by the Federal government are protected by the Privacy Act and the enabling statutes of some agencies.

In the European Union, informational privacy is guaranteed by national laws transposing the broad EU Data Protection Directive.[4] An example of such a law is the U.K. Data Protection Act.

Beyond conforming to law and ethical strictures, of course, the research community must always try to "do the right thing," both for reasons of basic decency and to earn and maintain the trust of members of the public who voluntarily allow themselves, data about themselves, or biospecimens from themselves to be studied in research.

The challenge is to protect privacy and foster research at the same time. A principal strategy for achieving this, especially in database-centered research, is to shield the identities of the people the data are about by blurring, removing, destroying or otherwise altering information that could lead to identification of the subjects. This kind of research is about cases and categories, not people.

---

[1] Federal Policy for the Protection of Human Subjects, HHS version (revised June 23, 2005); http://www.hhs.gov/ohrp/humansubjects/guidance/45cfr46.htm.

[2] HHS, Standards for Privacy of Individually Identifiable Health Information; www.hhs.gov/ocr/hipaa. For interpretation see "Protecting personal health information in research: understanding the HIPAA Privacy Rule," http://privacyruleandresearch.nih.gov/pr_02.asp.

[3] Regarding overlap between the two Rules, see Congressional Research Service, report LC(X):RL32909, "Federal protection for human research subjects: an analysis of the Common Rule and its interactions with FDA regulations and the HIPAA privacy rule" (updated June 2, 2005); among other websites, posted at www.fas.org/sgp/crs/misc/RL32909.pdf.

[4] Portal to the EU Data Protection Directive, the national laws, and the authorities who administer them: http://ec.europa.eu/justice_home/fsj/privacy.

Re-think open release of data?

Part of the backdrop to this project is the cultural habit of rapid, open release, or at least fairly open release, of genomic data that has been adopted by the involved scientists and institutions since the beginning of the genomic endeavor. Sociologically it is one of the defining features of the field. It has greatly facilitated research across political, sectoral, and disciplinary boundaries, and it has been consonant with, indeed has stimulated, the growing emphasis on scientific data-sharing and open publication generally.

As early as 1991 the National Center for Human Genome Research (the predecessor to NHGRI) and the Department of Energy required that sequence data be released within six months. Then in 1996 the International Human Genome Sequencing Consortium adopted so-called "Bermuda Principles," which encouraged rapid release into the public domain of sequence assemblies of 1-2 kilobases or greater.[5] In 1997 NHGRI required grantees to release sequence assemblies within 24 hours, and in 2000 it extended the policy to require weekly publication of raw sequence traces.[6] NIH and other institutions set up the necessary databases to receive and distribute the data. The approach served the fast-moving research so effectively that in 2003 the community developed more detailed "Fort Lauderdale Principles," outlining roles for funding organizations, data producers, and data users.[7]

HapMap is an example of a project that follows the Principles:[8]

> As is now standard practice in large-scale genomic research projects, the International HapMap Consortium follows a policy of releasing data as quickly as possible, anticipating that they will be useful for many investigators. The Consortium anticipates that the Project's data will be used in many ways, such as in developing new analytical methods, in understanding patterns of polymorphism, linkage disequilibrium, and haplotype associations, and in guiding selection of markers to map genes affecting specific diseases. Thus, the Consortium recognizes that the data are available to all users for any purpose.

NIH generally believes that "data should be made as widely and freely available as possible while safeguarding the privacy of participants and protecting confidential and proprietary data." Accordingly, it requires that applicants for grants exceeding $500,000 include in their application a plan for sharing final research data, or a clear justification of not sharing.[9] NIH's rationale is that:

---

[5] David R. Bentley, "Genomic sequence information should be released immediately and freely into the public domain," *Science* 274, 533-534 (1996).

[6] NHGRI, "Reaffirmation and extension of NHGRI rapid data release policies" (February 2003); www.genome.gov/10506537.

[7] Wellcome Trust (writing as convenor), "Sharing data from large-scale biological research projects: A system of tripartite responsibility" (2003); www.wellcome.ac.uk/assets/wtd003207.pdf.

[8] "The responsible use and publication of HapMap data"; www.hapmap.org/guidelines_hapmap_data.html.en.

[9] NIH, "Final NIH statement on sharing research data" (2003) and related documents; http://grants.nih.gov/grants/policy/data_sharing.

> Sharing data reinforces open scientific inquiry, encourages diversity of analysis and opinion, promotes new research, makes possible the testing of new or alternative hypotheses and methods of analysis, supports studies on data collection methods and measurement, facilitates the education of new researchers, enables the exploration of topics not envisioned by the initial investigators, and permits the creation of new datasets when data from multiple sources are combined.

(Less often mentioned as a rationale, at least until recently, is that early open release of data establishes their precedence as obvious "prior knowledge," thereby preventing their being claimed as proprietary know-how in patents and made less accessible for research use.)

There is no question about the research advantages of such principles or policies. Nor is there question about flexibility, as the Fort Lauderdale Principles and the policies based on them have never been construed as being absolute or encouraging the transgression of people's rights.

However. Part of the remit of this project is to examine whether in the future, genomic data, with various clinical or other associated data, will have to be modified to reduce identifiability and/or held back for controlled release, to a greater extent than has been done up to now.

A cautionary remark about language. Experts in this arena often speak of "public" data or "open" publication. But the usages are sloppy, and may refer either to data that truly are in the public domain, as when posted on a freely accessible website, or, quite differently, to data that only the "professional public" outside the custodian organization may apply to use, under conditional terms. Care should be taken with "public" and "open," as either can justifiably be understood by lay audiences as implying that data are being exposed to plain public view.

Issues arising as genomics matures

Many big new initiatives − including the Genetic Association Information Network (GAIN), The Cancer Genome Atlas (TCGA), the Genes and Environment Initiative (GEI), U.K. Biobank, the NHGRI Medical Sequencing Program, and many other projects − will:[10]

- □ generate data that are fine-grained, of wide genome coverage, and person-specific
- □ categorize many data with respect to disease-related genes or disease diagnosis
- □ maintain links, at least indirectly, to clinical, family, social, and demographic data
- □ and do all this on material from very large numbers of people.

[10] For brief sketches of a few such projects, see the Appendix.

And they will release the resulting data into a world that will:

- □ continue to assemble identified, or at least circumstantially characterized, police, military, and other DNA and genomic reference collections
- □ increasingly integrate genomic data with personal medical records
- □ as genotyping costs drop and knowledge increases, become more routinely capable of matching data to reference collections and inferring probabilistic implications for physical appearance, mental health, and illness risks
- □ continue to amass databases on most aspects of people's lives, with incentives to link those databases with genomic data for research, healthcare, public health, administrative, marketing, forensic, and other purposes
- □ continue to worry about the risks of erroneous or malicious identity disclosure and consequent embarrassment, blackmail, group stigmatization, financial fraud, or negative discrimination for health or life insurance, employment, promotion, mortgages, or loans.

Recently several observers have served serious notice that genomic data are becoming more identifiable.

Malin and Sweeney showed that DNA sequences unlabelled as to demographics or identifiers, if interpreted for some common disease genes and probabilistically screened against publicly available data (such as the detailed hospital discharge data that are publicly accessible in some States), can sometimes be narrowed-down to a few individuals.[11]

In a different approach, Malin argued that DNA sequences can be mapped against family disease-incidence patterns, hospital visit patterns, and other data, and be identified by "trail analysis."[12]

Concerned that "genome-wide association studies now routinely use more than 100,000 SNPs to genotype individuals" and that current protections are inadequate, McGuire and Gibbs have recommended that sequencing research be clearly designated human-subjects research − thus requiring more elaborate consent and closer scrutiny by Institutional Review Boards (IRBs) − and that tiered approaches, in which the data-subjects have more say in determining how data are released and by whom they can be used, should be adopted for release of genomic data.[13]

---

[11] Bradley Malin and Latanya Sweeney, "Determining the identifiability of DNA database entries," *Proceedings of the American Medical Informatics Association Symposium 2000*, 537-541 (2000); available at http://privacy.cs.cmu.edu/dataprivacy/projects/genetic/dna1.html.

[12] Bradley A. Malin, "An evaluation of the current state of genomic data privacy protection technology and a roadmap for the future," *Journal of the American Medical Informatics Association* 12, 28-34 (2005).

[13] Amy L. McGuire and Richard A. Gibbs, "No longer de-identified," *Science* 312, 370-371 (2006).

Getting things right in this area, Foster and Sharp have remarked, will have implications broader than (just) facilitating genomic research:[14]

> The challenges of using linked genotype−phenotype data for medical-sequencing projects prefigures issues that will arise in future uses of many existing biological samples linked to phenotypic information, including disease registries, hospital-based tissue collections and prospective cohort studies. Thus, developing effective strategies for addressing these challenges in medical-sequencing research can inform a much broader set of issues in the ethical conduct of research.

And finally, an observation by this author about subject selection and a question about consent. The Human Genome and HapMap projects have genotyped DNA from only a few very carefully selected people who have consented to the analysis and open publication only after thorough explanation and discussion. But such painstaking selection and consent negotiation cannot as a general matter be expected for future projects involving many people. As an ethical matter, should consent be relied upon to justify open publication of data that are potentially identifiable?

There is ample cause for concern, and work to be done.

---

[14] Morris W. Foster and Richard R. Sharp, "Ethical issues in medical-sequencing research: implications of genotype−phenotype studies for individuals and populations," *Human Molecular Genetics* 15, R45-R49 (2006).

2. Identifiability and identifiers

"Identifiability"

Identifiability is the potential associability of data with persons. Identifiability runs a spectrum, from overtly identified, to possibly deductively identifiable, to absolutely unidentifiable.

In legal regimens, indirect identifiability is as important as direct. For instance, the HIPAA Privacy Rule applies to "individually identifiable health information," i.e. "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" (§160.103).

Similarly, the U.K. Data Protection Act applies to all "personal data," "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," the person legally responsible for determining the purposes for which, and manner in which, the data are handled (§I.1-(1)).[15]

If data aren't identifiable they aren't "personal," and a variety of rights and obligations that apply to personal data may not be relevant. General rights of informational privacy, and professional obligations of medical confidentiality, for example, usually apply only to data that are "about" real people.

Almost all health data are initially collected as identified data, whether for healthcare, public health, or research purposes. Data can be de-identified in a variety of ways and to varying degrees, either irreversibly or reversibly. De-identification is a crucial strategy for research.

Identifiers

For any set of data about people, three sorts of identifying factors − commonly, although a bit too casually, referred to as "identifiers" − can be distinguished:

> **Administrative or demographic tags** (name, Social Security number, email address, hospital name, Zip code...)
>
> **Overt descriptors** (gender, eye color, height, blood type, scars, asthma...)
>
> **Indirect clues** (medication use, number of children, spouse's occupation, circumstances of emergency-room admission...).

---

[15] Although the Act applies only to data about living individuals, professional guidance in the U.K. advises that medical data should be held in confidence after death as well. Because of the implications for relatives, the issue of protection of DNA and genomic data after death warrants re-evaluation everywhere now.

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. The Rule provides that for data to be considered adequately de-identified and therefore not subject to its provisions, a number of listed descriptors must be absent. (See Figure 2, at the end of this chapter, known to aficionados as "the HIPAA List.").

The List comprises identifiers that are linked fairly directly somewhere to name-and-address; it does not include all prime descriptors of persons. For example, there is no element on the list for health, illness, or disability characteristics, even those that may be evident to simple perception such as hearing impairment, palsy, albinism, limp, or wheelchair dependency. Presumably the assumption is that these will be caught by the "any other" element (R), even though this relegates a lot to judgment and the qualifier "unique" is subject to interpretation. One may wonder whether (R) covers, for example, an International Classification of Diseases code ("ICD-10 L40" = psoriasis vulgaris).

Knowing a few of the 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 through some other source, such as a Social Security look-up database. It is obvious that some identifying elements are pretty weak in evidentiary power, at least if they aren't linked with other data. But some others, such as birthdate, are stronger.

A word about Limited Data Sets, since they are one of HIPAA's concessions to research and are being used in some genomic projects. For research, the Privacy Rule allows data custodians to release Limited Data Sets, data from which many but not all core identifiers have been stripped (§164.514(b)(5)(e)(2)). Names, electronic communication addresses, and biometric identifiers must not be present, for example, but gender, birthdate, treatment dates, cities, states, Zip codes, and some other potentially identifying clues can remain, as well as substantive health information. When applying to a data custodian to use a Limited Data Set, researchers must specify which data they want and the intended uses, name who will be using the data, commit to enforcing safeguards, and promise that they will not attempt to identify the data-subjects or contact them.

Caution regarding the identifiability of Limited Data Sets is expressed in policies such as that of the Centers for Medicare & Medicaid Services:[16]

> Limited Data Sets (LDS) contain beneficiary level health information but exclude specified direct identifiers as outlined in the Privacy Rule. LDS are considered identifiable even without the specified direct identifiers. Because the information is considered identifiable, it remains subject to the Privacy Act of 1974 as well. These data are identifiable because of the potential for identifying a beneficiary due to technology, particularly in linking and re-identifying data files.

---

[16] "Procedures for Limited Data Sets," www.cms.hhs.gov/PrivProtectedData/10_LimitedDataSets.asp. Provision of CMS data is governed by strict Data User Agreements.

Because genomic data have exquisitely fine-grained informational structure, describe fundamental constituents of the person's body, don't change during the lifetime, and can be used for matching and possibly for profiling, surely they must be treated as strong identifiers. This has implications for the safeguarding of genomic data and the ethico-legal standards that govern that safeguarding. For present purposes it is important to contemplate whether now or in the future the HIPAA identifier elements (P), "Biometric identifiers," or (R), "Any other unique identifying number, characteristic, or code," should be interpreted as including genetic or genomic information.

This chapter has used the HIPAA Privacy Rule to illustrate the points, but similar issues arise with all privacy protection regimes.

Terminology

A note regarding terms for various states of identifiability. This author prefers to speak of data as being, simply, either identified or identifiable, or key-coded, or non-identifiable.[17] Approximate synonyms used in various professional cultures are as follows.

Figure 1. Concordance of identifiability terms

identified or identifiable	**key-coded**	**non-identifiable**
personal nominative	reversibly de-identified linked anonymized pseudonymized pseudoanonymized encrypted coded	irreversibly de-identified unlinked anonymized unidentifiable anonymous

Use of "key-coded" avoids such awkward expressions as "pseudonymized" and helps the public understand the approach. "Encryption" is now taken in everyday speech to mean the scrambling of messages to keep them secret en route. "Coding" is universally used in the health sciences to refer to the classification of diseases, drugs, and procedures to standard categories. The central feature of a system that maintains the potential to reassociate substantive data with identifying data is the key: hence, key-code.

[17] For some considerations in key-coding see William W. Lowrance, *Learning from Experience: Privacy and the Secondary Use of Data in Health Research* (The Nuffield Trust, London, November 2002), pp. 32-33; www.nuffieldtrust.org.uk/ecomm/files/161202learning.pdf

Figure 2. The HIPAA Privacy Rule's identifier list (§164.514(b)(2))

(i) A covered entity may determine that health information is not individually identifiable health information only if... the following identifiers of the individual or of relatives, employers, or household members of the individual, are removed:

(A) Names;

(B) All geographic subdivisions smaller than a State, including street address, city, county, precinct, zip code, and their equivalent geocodes, except for the initial three digits of a zip code if, according to the current publicly available data from the Bureau of the Census:

(1) The geographic unit formed by combining all zip codes with the same three initial digits contains more than 20,000 people; and (2) the initial three digits of a zip code for all such geographic units containing 20,000 or fewer people is changed to 000;

(C) All elements of dates (except year) for dates directly related to an individual, including birth date, admission date, discharge date, date of death; and all ages over 89 and all elements of dates (including year) indicative of such age, except that such ages and elements may be aggregated into a single category of age 90 or older;

(D) Telephone numbers;

(E) Fax numbers;

(F) Electronic mail addresses;

(G) Social security numbers;

(H) Medical record numbers;

(I) Health plan beneficiary numbers;

(J) Account numbers;

(K) Certificate/license numbers;

(L) Vehicle identifiers and serial numbers, including license plate numbers;

(M) Device identifiers and serial numbers;

(N) Web Universal Resource Locators (URLs);

(O) Internet Protocol (IP) address numbers;

(P) Biometric identifiers, including finger and voice prints;

(Q) Full face photographic images and any comparable images; and

(R) Any other unique identifying number, characteristic, or code...

[and]

(ii) The covered entity does not have actual knowledge that the information could be used alone or in combination with other information to identify an individual who is a subject of the information.

3. *Strategies for identifying non-identified genomic data*

With detailed individual-level genomic data, such as SNPs and sequences, there are three basic and very different approaches to identifying the person to whom the data pertain, which will be discussed in turn:

- Matching genotype against reference genotype
- Linking genomic+associated data with other data
- Profiling from genomic characteristics.

These strategies can be used for socially desired purposes such as fighting crime and terrorism and identifying victims of accidents, disasters, epidemics, or war. They can also be used for nefarious purposes.

Matching genotype against reference genotype

Lin, Owen, and Altman have analyzed the chances of random matching of unidentified genotype data against reference-collection data and concluded that "specifying DNA sequence at only 30 to 80 statistically independent SNP positions will uniquely identify a single person." (They mean uniquely *match*, i.e. confirm that two samples come from the same identical person; whether it *identifies* anybody in the normal-talk sense depends on whether the reference data themselves are personally identified.) They go on and say that randomly changing 10% of SNPs, or binning using standard statistical techniques, do not change this conclusion.[18]

Identified forensic-purpose biospecimens from millions of people are held by the criminal justice system and the armed services. Most forensic matching focuses on a standard panel of short tandem repeat polymorphisms ("STRPs"). The identification efficacy is as high as it would be if SNPs were used.[19]

Literally countless biospecimens, and a growing number of genomic analyses, are held by medical, public health, and health research institutions. A practical source of reference for identifying victims and criminals is blood relatives' genotypes.[20]

Matching is possible to a high degree of certainty if very much of the genome is available. Its reliability is substantially higher than that with the matching of fingerprints or retinal scans.

---

[18] Zhen Lin, Art B. Owen, and Russ B. Altman, "Genomic research and human subject privacy," *Science* 303, 183 (2004), with supporting calculations at www.sciencemag.org/cgi/content/full/305/5681/183/DC1.

[19] A standard text is John M. Butler, *Forensic DNA Typing*, 2nd edition (Elsevier, Amsterdam and Boston, 2005).

[20] Frederick K. Bieber, Charles H. Brenner, and David Lazar, "Finding criminals through DNA of their relatives," *Science* 312, 1315-1316 (2006).

Questions about matching include:

Q1: How "much" genome, i.e. how many megabases, SNPs, STRPs, traces, genes, or other amounts of information, is sufficient for identifying by matching? Surely a gene or two is not enough, but how much is?

(The answer seems to be, "It depends" − it depends on the density, or resolution, of mapping, the extent of genome covered, the rarity of variants, the degree of linkage disequilibrium, and other factors.)[21]

Q2: Is it possible to draft at least semi-quantitative criteria for setting thresholds of identifiability, for instance in technical guidance?

(The answer to the previous question implies that this will remain a matter of judgment. But maybe risk-analytical approaches can be devised.)

Linking genotype+associated data with other data

A second route to identifying genotyped subjects is deduction by linking and matching genotype+phenotype (or +other data) with data in health, demographic, administrative, employment, criminal, military service, hazard exposure, disaster response, or other databases. Often the context as regards exposure, disease, or locale strongly suggests which external databases may yield useful information. If the data linked-to are overtly identified, the task is straightforward. If the data linked-to are not fully identified, inferential matching and narrowing-down may be possible.

Statisticians possess an array of well tested techniques for identifying data-subjects from partial data, and an equally well tested array of techniques for obfuscating inferential identification.[22, 23]

There is no shortage of external sources of identified data. There are so many public and commercial databases about people's lives, especially in the U.S., that it requires databases of databases and enables a lucrative look-up/hunt-down industry.[24] Just a few familiar examples of accessible data are those about birth, marriage, divorce, and death, home and business addresses and telephone numbers, voter registration, motor vehicle and boat registration, real estate ownership, police and court proceedings, organization membership,

[21] Zhen Lin, Russ B. Altman, and Art B. Owen, Letter, "Confidentiality in genome research," *Science* 313, 441-442 (2006).

[22] A rich source is the American Statistical Association's website on Privacy, Confidentiality, and Data Security; www.amstat.org/comm/CmtePC. Another is the Federal Committee on Statistical Methodology's *Working Paper 22*; "Report on statistical disclosure limitation methodology," www.fcsm.gov/working-papers/spwp22.html.

[23] A standard text is Josep Domingo-Ferrer, editor, *Inference Control in Statistical Databases* (Springer-Verlag, Berlin, 2002).

[24] Thousands of databases, many of them holding voluminous personally identified information, can be consulted via services such as www.searchsystems.net and www.choicepoint.com.

professional licensing, government employment, and in some states hospital discharge. Family pedigree databases can provide complementary information.

Beyond these databases, of course the health arena holds uncountably more confidential but conditionally accessible data ranging from medical care and payment records to perinatal genetic screening results, Guthrie cards, disease registries, and implant registries.

Obviously, barring the availability of a reference genotype collection, genotype+phenotype (or +other) data are much more vulnerable to being inferentially identified than genotype data alone are. Thus careful attention will have to be paid as health research moves, inevitably, toward linking genomic data with clinical and social data.

Profiling from genomic characteristics

Increasingly now there is concern about whether a probabilistic profile of an individual can be inferred from genotype. This amounts to *describing* rather than actually *identifying*. Making such inferences depends on being able to interpret how particular genomic factors contribute to determining bodily characteristics or behavioral or disease likelihoods, and then developing a composite description of the person. Such a description can only be a probabilistic profile against which candidates can be screened, and any further narrowing-down depends on linking with other evidence.

As the population frequencies of SNPs and STRPs become better known, both kinds of markers are being used to construct profiles of ethnicity. Forensic approaches have tended to use STRPs, perhaps because STRP data are the principal sort held in police collections (for matching), but profiling based on them suffers from such limitations as the fact that STR loci mutate relatively rapidly.[25] As millions of SNPs become analyzed and the ancestry of the DNA sources is characterized, the patterns tend to suggest profiles.[26]

It is sometimes rumored that the FBI, CIA, and other agencies are developing systems for profiling suspects based on genomic data, and not just with reference to ethnicity. Apparently from time to time they have explored this avenue, but so far have not found it very productive.[27] They should be expected to pursue genomic profiling when the science has advanced sufficiently.

[25] For a review see John M. Butler, "Genetics and genomics of core short tandem repeat loci used in human identity testing," *Journal of Forensic Sciences* 51, 253-265 (2006); ethnicity profiling is discussed on p. 260.

[26] David A. Hinds, Laura L. Stuve, Geoffrey B. Nilsen, Eran Halperin, Eleazar Eskin, Dennis G. Ballinger, Kelly A. Frazer, and David R. Cox, "Whole genome patterns of common DNA variation in three human populations," *Science* 307, 1072-1079 (2005).

[27] Anecdotal communications to the author.

Again: Any inference derived by comparing an individual's SNP or STRP markers against the prevalence of those markers in culturally or geographically defined populations can at best yield only a likelihood ratio.[28, 29, 30]

Questions about profiling include:

> **Q3**: How accurate can profiling be regarding possible ethnic/racial origins or appearance (acknowledging the definitional ambiguities)?
>
> **Q4**: Which corporal features can be inferred now, apart from gender, the odds on blood type, skin pigmentation, and overt manifestations of Mendelian disorders? Which might be expected to be deducible before long − height, shoulder width, or other aspects of skeletal build? Hair color or texture? Eye color? Eye shape or other facial features? Cranial, dental? Others?
>
> **Q5**: Which behavioral or disease attributes − likelihood of being depressive, schizophrenic, alcoholic, violent? Diabetic, hypertensive? Others?
>
> **Q6**: Shouldn't we assume that in 5−10 years many attributes will be profilable?
>
> **Q7**: With profiling, "how much" genome has to be known before the data in themselves are usefully descriptive? Again, is it possible to set semi-quantitative thresholds of identifiability?
>
> (The answer will be very different from that to the same question regarding matching.)

---

28 For perspective see Noah A. Rosenberg, Jonathan K. Pritchard, James L. Weber, Howard M. Cann, Kenneth K. Kidd, Lev A. Zhivotovsky, and Marcus W. Feldman, "Genetic structure of human populations," *Science* 298, 2381-2385 (2002).

29 NHGRI Race, Ethnicity, and Genetics Working Group, "The use of racial, ethnic, and ancestral categories in human genetics research," *American Journal of Human Genetics 77*, 519-532 (2005).

30 Susanne B. Haga, "Policy implications of defining race and more by genome profiling," *Genomics, Society and Policy* [online] 2 (1), 57-71 (2006); www.gspjournal.com.

4. Strategies for de-identifying genomic data

For reducing identifiability of genomic (perhaps +other) data before releasing them for research, there are three sorts of technical options, which will be discussed in turn:

- □ Limiting the proportion of genome released
- □ Statistically degrading the data before releasing
- □ Sequestering identifiers via key-coding.

Limiting the proportion of genome released

The first option is to publish only limited segments of genomes, such as sequence traces or only a few variants, along with minimum necessary phenotypic or other data. This requires judgment as to how much to release, which will depend on what genomic region is involved and the circumstances of data or biospecimen collection (for instance, whether the data are openly known to be about people having a particular disease, or who live in a certain region), and be difficult to generalize. It may deny important data to researchers, including data about regions of the genome that they can't know whether they need to know.

In practice many disease-specific projects do limit the portion of genome they release, but it is not clear that they use any criteria other than "no more than necessary." How such a limitation might apply with whole genome association studies, however, is unclear [at least to this author]. Certainly, precautions can be taken, such as releasing sequence traces in such a separated manner that it isn't possible for an external analyst to reconstruct which traces pertain to a single individual's DNA.

The pivotal issue here is the same as was raised in the previous chapter: how to know "how much" genome is too much to release.

Statistically degrading the data before releasing

The second option is to degrade data before posting, such as reducing precision by lumping <u>G</u> and <u>A</u> as purines, and <u>C</u> and <u>T</u> as pyrimidines; or "fuzzing" data by adding statistical noise, i.e. spurious but not dissimilar data, to data-sets; or randomly altering or exchanging a small percentage of SNPs; or micro-aggregating or "binning" small subsets of data in various ways to reduce granularity. These are all standard statistical disclosure-reduction techniques, although they have to be carefully adapted for genomic data.[31, 32]

---

[31] Regarding such techniques generally see footnote 22 above.

[32] Statistical approaches to estimating and reducing the disclosure risks of SNP databases are explored in Zhen Lin, "Balancing utility and anonymity in public biomedical databases," doctoral dissertation, Stanford University (April 2005); http://helix-web.stanford.edu/people/zlin/pubs/zlin_thesis.pdf.

Data transformed in such ways may meet the needs of some query systems in which researchers pose questions to external databases. And they are fine for some higher level population surveys. But at the sequence level the human genome data-tape comprises some 3,000,000,000 data-cells − arrayed in linear order, though segmented into chromosomes − and the sensitivity of the human organism to variance is such that 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 of data simply degrades usefulness.

A pragmatic question on which it is very important to have calibration from genomicists is:

> **Q8**: How serious an impediment to research is masking, binning, perturbing, or otherwise degrading genomic data? How does the answer vary with different techniques, and with different lines of research?

Sequestering identifiers via key-coding

The method most widely used in health research for de-identifying data is key-coding (reversibly de-identifying), in which potentially identifying data are separated from the substantive data, such as health data, but a link is maintained by assigning an arbitrary code number to each part of the data–identifier pair before they are separated. Held securely and separately, the key allows reassociating the substantive data with the identifiers if that is ever necessary. The key and the responsibility for its use can be delegated to a trusted party, and use of the key can be guided by agreed criteria and subjected to oversight.

The scientific and ethical advantages of reversible de-identification are widely appreciated and need not be reviewed here. Key-coding can be used among multiple databases and with biospecimens, and it can keep elements, such as clinical data and biospecimens, cross-referenced with each other even if the links to the data-subjects are irreversibly severed. For high-sensitivity data, the codes can be further encoded. Elaborate key-coding systems and identifiability vocabulary are being developed for pharmacogenetic data submitted in regulation.[33, 34]

Surely what is important in any instance is not whether a link of some sort exists somewhere, but *whether the identifiers can be known to the researchers* who study the substantive data. A carefully constructed key system can provide reliable safeguards. Furthermore, data-use agreements almost always require that data recipients promise not to attempt to re-identify the data-subjects, which obviously forbids abusing the key system.

[33] European Medicines Evaluation Agency, "Position paper on terminology in pharmacogenetics" (2002); www.emea.eu.int/pdfs/human/press/pp/307001en.pdf.

[34] Standardization of vocabulary is now being considered by the International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use (ICH).

The HHS Office for Human Research Protection (OHRP) has issued helpful guidance on these issues. The document should be consulted for details, but a main point is this:[35]

> OHRP considers private information or specimens not to be individually identifiable when they cannot be linked to specific individuals by the investigator(s) either directly or indirectly through coding systems. For example, OHRP does not consider research involving *only* coded private information or specimens to involve human subjects as defined under 45 CFR 46.102(f) [of the Common Rule] if the following conditions are both met:
>
> > (1) the private information or specimens were not collected specifically for the currently proposed research project through an interaction or intervention with living individuals; and
> >
> > (2) the investigator(s) cannot readily ascertain the identity of the individual(s) to whom the coded private information or specimens pertain because, for example:
> >
> > > (a) the key to decipher the code is destroyed before the research begins;
> > >
> > > (b) the investigators and the holder of the key enter into an agreement prohibiting the release of the key to the investigators under any circumstances, until the individuals are deceased...;
> > >
> > > (c) there are IRB-approved written policies and operating procedures for a repository or data management center that prohibit the release of the key to the investigators under any circumstances, until the individuals are deceased; or
> > >
> > > (d) there are other legal requirements prohibiting the release of the key to the investigators, until the individuals are deceased.

NIH and other investigators have had considerable experience with such key-coding, but the craft continues to deserve refinement. The problem is not one of technology development − many identifiability-protecting encryption programs exist, and ever more sophisticated ones are being developed (for protecting confidentiality while mining data in large sets electronic medical records, for example) − but one of discipline in day-to-day practice.

Key-coding can effectively sequester obvious identifiers. But before data are released, various indirect identifying bits (such as free-text narrative or mentions of family members) may have to be stripped off, or if these data are absolutely necessary for the research, protected by a non-disclosure agreement. Artificial-intelligence programs can help, for example by screening for proper nouns or especially sensitive words or phrases, but the exercising of human judgment cannot be avoided.

---

[35] HHS, Office for Human Research Protection, "Guidance on research involving coded private information or biological specimens" (2004); www.hhs.gov/ohrp/humansubjects/guidance/cdebiol.pdf

Obviously the risk questions raised by this chapter are the same as those of the previous chapter, although coming from the opposite direction − essentially, “How much” genome is too much to hang out in public?

Related to de-identification are several questions of responsibility:

> **Q9**: Who should be responsible for de-identifying data before providing them, via whatever route, for research? Are physicians and principal investigators adequately prepared to do this? What roles should IRBs play with respect to disclosure control? [36]

> **Q10**: For data or biospecimens provided to research resource platforms, such as GAIN or UK Biobank, should the platforms themselves conduct any identifiability or disclosure review before releasing the data onward, via whatever route, for research?

[36] Regarding IRB issues see Virginia de Wolf, Joan E. Silber, Philip M. Steel, and Alvan O. Zarate, “Meeting the challenge when data sharing is required,” *IRB: Ethics & Human Research* 28 (2), 10-15 (2006).

5. Controlled release

The alternative to open publication is release of data to researchers under agreements that, inter alia, protect privacy and confidentiality. Legally such agreements amount to contractual undertakings. In many instances they are also public promises, as when they pertain to data obtained from government institutions or sources supported by government funds. Often for legal clarity the definitions or conditions in agreements refer to definitions or requirements of the Common Rule, the HIPAA Privacy Rule, the Privacy Act, or other statutes, regulations, or guidance.

Many projects release data through two precautionary stages: first reducing identifiability to a reasonable extent by suppressing overt identifiers, broadening data (such as rounding birthdate to year of birth or age-range), and so on, and performing at least informal disclosure review; and then providing access under a controlled-release agreement.

Terms of agreements

The terms commonly addressed by controlled-release agreements are shown in Figure 3 (at the end of this chapter), meant here to telegraph the considerations. Probably no agreement anywhere includes all of these terms, and some agreements may incorporate additional ones, but the table is a basic menu. Among the terms most relevant for protection of identifiability are the following.

Consent relating to identifiability. May address the honoring of commitments to protect identifiability in various ways. May set limitations on purpose, allowed users, or other aspects of data use, either for the whole data-set or for particular data-subjects, which may reflect perceptions of trustworthiness. May address publication, and either promise that identities will be thoroughly obscured or warn that some non-negligible identifiability risk may accompany publication.

Confidentiality protection. Must always include physical, organizational, and IT security. May make reference to compliance with International Standards Organization, HIPAA, or other security standards. Usually specifies who, if anyone, will be responsible for de-identifying data and the procedures and criteria for doing this. May cover processes of key-coding, safeguarding of the key, and use of the key. Almost always states that researchers will make no attempt to re-identify or recontact the data-subjects.

Limiting of onward transfer. Restricts transfer (or access, which amounts to the same thing) and extends the chain of confidentiality and the accompanying obligations.

Linking. May be discussed if, inter alia, linking is contemplated that might have the effect of increasing identifiability.

Publication, release, and returning of findings. Specifies whether data must be fed back to the original data resource, and whether publication of detailed findings, such as sequences, may be fully open, as on a publicly accessible database, or must be by further controlled release. Usually requires that the identifiability of any derived data must be protected to the same extent as that of the data being provided.

IRB or other ethics approval. Deferred to for oversight regarding conformance to the Common Rule and other obligations. An issue may be at what stage(s) IRB review should be carried out.

Arrangements for controlled release

Many details of practicality and governance have to be attended to, but data-providers and data-users are generally familiar with controlled-release arrangements. These three examples provide a few specifics.

- Wellcome Trust Case Control Consortium. The Consortium, which involves a number of university units in the U.K., is genotyping de-identified DNA samples from thousands of people with known chronic diseases, and controls. Access to cleaned, raw, and summary data is decided by a Consortium Data Access Committee and is subject to a Data Access Agreement.[37]

- Framingham Heart Study. Framingham releases DNA and data via evaluation by two committees and a Data and Materials Distribution Agreement.[38]

- Genetic Association Information Network (GAIN). Because it will genotype DNA submitted by a network of contributing academic disease-specific studies having diverse auspices, consents, and IRB stances, and because it is concerned about identifiability of the extensive data it expects to generate, the GAIN project recently changed its policy from one envisioning fairly open publication to one using controlled release. De-identification will primarily be the responsibility of the data providers. Access will be controlled by a Data Access Committee and will be subject to Data Access Certification.[39]

The GAIN example illustrates the importance for the genomic research community of exploring criteria for deciding whether particular sorts and “amounts” of data can be posted publicly or must be managed by controlled access.

---

[37] www.wtccc.org.uk.

[38] www.nhlbi.nih.gov/about/framingham/policies/index.htm.

[39] www.fnih.org/GAIN/Updated_Data_Access_Policy.shtml.

Alternatively to transferring full data-sets to external researchers, more restricted channels can be employed, such as:

- □ on-site data enclaves (or Research Data Centers, as several Federal ones are called) to which researchers come and perform studies on a database in a secure, dedicated, monitored server
- □ remote-query systems, in which researchers interrogate databases and obtain responses, possibly veiled in some ways, via secure telecommunications
- □ service analyses that analyze data or biospecimens according to agreed methodology and provide the results to the commissioning researchers.

Ways must be devised to make controlled release practical, binding, and palatable − conditions that are not procedurally onerous but that at the same time secure genuine, formal, enforceable commitments. Perhaps for some group of projects a general data-use license can be worked out, for instance, through which researchers, with their institutions, agree to terms and gain entrée to a large suite of data and/or biospecimens.

> **Q11**: Is there any appeal in exploring broad data-use licenses for access to centrally held genomic data? Has there been any relevant experience with such a scheme?

Figure 3. Some terms of data and biospecimen release agreements [40]

Screening of scientific competence and merit	(why screen − to protect the resource? to conserve effort?) purposes, potential scientific payoff...
Specification of what is provided	data, biospecimens, analysis, informatics, linking, assistance, training, archiving?
Consent	coverage, documented where, tracked how...
Purpose limitation	disease-specific use? start cell lines? commercial use?
Confidentiality	de-identify (how)?, promise not to try to re-identify, security, training, ...
IRB or other ethics approval	...at point of data collection? at research resource platform stage before release?
Limiting of onward transfer	restrictions
Linking	expectations, restrictions
Recontacting data-subjects	justifications, recontact by whom and how
Maintaining quality of resource	promise to deal with errors or contamination
Publication or returning of findings	required? publish how? timing?
Acknowledgments	“much obliged”
Co-authoring	required for control or credit-sharing?
Enriching the resource	integrate findings into the resource? who is responsible for quality?
Informing data-subjects	...of progress? of person-specific findings?
Archiving	how? who pays? conditions of access?
Intellectual property rights	IP assignments or waiving
Responding if a subject withdraws	destroy data, biospecimens, or links?
Returning or destroying materials	...when finished? if commitments are broken?
Prioritization of access to data or biospecimens	...if biospecimen quantity or analytic or IT resources are limited
Fees	for what? does fee depend on IP prospects?
Transborder enforcement	legal constraints, ethics approval, subjects’ rights
Monitoring, oversight, or audit	plans, expectations
Contingencies if resource or project elements are terminated	destroy the resources? pass on to another institution that will preserve the conditions?
Legal disclaimers	[not responsible for quality or consequences]

[40] Adapted from William W. Lowrance, “Access to Collections of Data and Materials for Health Research” (March 2006); http://www.wellcome.ac.uk/doc_WTX030843.html.

6. Identifiability risks, overall

All of this must be examined from risk perspectives − risks to data-subjects, risks to data stewards, risks to researchers and their institutions, even risks to the genomic research enterprise. Concern must be about whether data can be used to (a) deduce the identity of data-subjects or (b) deduce facts about data-subjects, and whether in either case this can lead to harm.

Sizing-up risks of any kind involves two activities. First, "risk assessment" estimates the probability of undesired events compounded by the severity of the likely consequences. Then, "risk appraisal" weighs the risks in perspective of personal or societal values. Appraisal can be cast as willingness to invest in reducing the risk by reducing the odds or the stakes or both. A broader appraisal can weigh the risks against benefits gained in the risktaking and consider whether the risks are acceptable. This is the way people think about most situations in life, whether they realize it or not.[41]

Genomic disclosure-risk-assessment must take account of such factors as the extent of genome covered; the density, or resolution, of mapping; the rarity of variants (because rarity increases identifiability); the degree of linkage disequilibrium; and the specificity with which gene effects are known. A special consideration with genomics is the disclosure risk for blood-relatives of people whose genome is studied, which has implications for consent and for safeguards.

Safeguards can't be discussed here except to state the sermonic point that an array of physical, procedural, cybersecurity, training, and legal safeguards must be in place against both accidental release and intrusive access. Among other reason*s, safeguards are what justify asking for broad consent.*

(A large topic that must be left to other forums is the need for genetic anti-discrimination laws, which tend to focus more on harmful consequences than on processes.)

What are the threats? We know that computerized systems can be broken into, data obtained by subterfuge, laptops stolen, and biospecimens transferred improperly. To the present there have been remarkably few proven abuses of medical data, much less health research data. But with the coming of electronic medical records, increased linking of databases, and so on − and given the vague foreboding that many people feel about anything "genomic" − public concerns are intensifying. As was mentioned at the outset of this document, the abuses that can be imagined range from embarrassment, blackmail, fraud, and group stigmatization, to negative discrimination for health or life insurance, employment, promotion, mortgages, or loans. Another possible abuse, depending on point of view, is unconsented parentage testing.

[41] Such odds−stakes thinking comes naturally when deciding where to cross a street, which ski run to attempt, whether to carry an umbrella (to the gym vs to a wedding), what intimate concerns to confide (to a close friend vs to a casual acquaintance), how candid to be in answering a questionnaire, whether to volunteer as a subject in alcoholism research....

Should we expect accidental releases, hacking, and attempts at abuse? Certainly. Detailed speculation is fruitless, although risk-anticipation exercises can help identify vulnerabilities and suggest defenses. It must be assumed that some threats are real possibilities, and that some can have serious consequences.

A plea for risk aversion. Surely it will be important not to expose "too much" of people's genomes in the coming years, only to regret it in the future when the analytic technologies become more robust, affordable, and routine, and genomic information becomes more easily abusable.

Q12: Overall, how much should we fret over genomic disclosure risks? (Details?)

Q13: Should the research community worry very much about access to protected genomic research data or biospecimens by the police, FBI, or other forces of public order, as compared with accidental release or malicious intrusion? Why? What, if anything, should it be doing differently?

Q14: Is there any reason to re-examine the Fort Lauderdale Principles, given that they are flexible and voluntary? Would any other forms of guidance be useful?

7. Flanking issues

Here, in no particular order, are some aspects of the larger puzzle that this project couldn't address but that very much need to be pursued.

Construral of genomic "human subject" under the Common Rule and other regulations. The fact that this is a perennial issue doesn't mean it shouldn't be worked on. Genomic research faces difficult questions regarding such matters as the status of people as "subjects" (or not) whose data or biospecimens have been assiduously de-identified, and the status as subjects of uninvolved relatives of people whose specimens are genotyped or being considered for genotyping. The answers have implications regarding, inter alia, identification and consent.

Consent. As an ethical matter, should consent be relied upon to justify deposition, in a publicly-accessible database, of data that have some realistic chance of being identifiable?

Controlled-release arrangements. As was suggested at the end of chapter 5, arrangements need to be explored that meet the ethical, legal, IT, managerial, and public perception challenges, and at the same time don't erect impractical barriers to research. Needing to be addressed with this are the special issues that arise when access to data is provided across national jurisdictions.

Certificates of Confidentiality, the legal assurances that NIH can issue under the Public Health Service Act 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."[42] They offer protection, but have limitations. How useful can the Certificates, or for that matter any conceivable legal ring-fence against forced disclosure, be for genomic projects?

Genetic anti-discrimination laws. The legislative saga rumbles on....

Protection of information on deceased persons. Continuing protection after death is required under the HIPAA Privacy Rule, but [bizarrely, to this observer], not under the Common Rule. The directness of the implications of people's genomic data for surviving relatives makes this issue more important for genomics than for most other health sciences.

[42] NIH Certificates of Confidentiality Kiosk: http://grants.nih.gov/grants/policy/coc.

Appendix. Sketches of a few projects

Framingham Heart Study. A study, begun in 1948, of the causes of cardiovascular and related diseases that has followed a cohort of some 5,200 people originally living around Framingham, Massachusetts, and many of their children, and is now recruiting grand-children. In its latest phase Framingham has begun examining genetic factors. (www.framingham.com/heart, and www.nhlbi.nih.gov/about/framingham)

Genes and Environment Initiative (GEI). An ambitious proposed NIH-wide program to analyze genomic factors, develop improved technologies for monitoring exposures, and study how genes and exposures interact as risk factors of disease. While Congressional budget approval is pending, elaborate concept exploration is being conducted. (http://grants1.nih.gov/grants/guide/rfa-files/RFA-HG-06-033.html)

Genetic Association Information Network (GAIN). A public-private cooperative project of the Foundation for the NIH, NIH, Pfizer Inc, Affymetrix Inc., the Broad Institute, and Abbott Laboratories. Will perform whole genome association studies on samples provided from existing case-control studies of patients having common diseases. Full planning is underway, with initial funding decisions to be announced soon. (www.fnih.org/GAIN/GAIN_home.shtml)

Genome wide association study (GWAS). A generic term, which NIH defines as including "any study of genetic variation across the entire human genome that is designed to identify genetic associations with observable traits (such as blood pressure or weight), or the presence or absence of a disease or condition."[43]

National Health and Nutrition Examination Survey (NHANES). A long-running series of examination surveys conducted by the Centers for Disease Control and Prevention (CDC). NHANES makes DNA from many later-phase cohort members available for analysis under very tightly controlled conditions, but it does not allow release of any genomic data. (www.cdc.gov/nchs/about/major/nhanes/research_proposal_guidelines.htm)

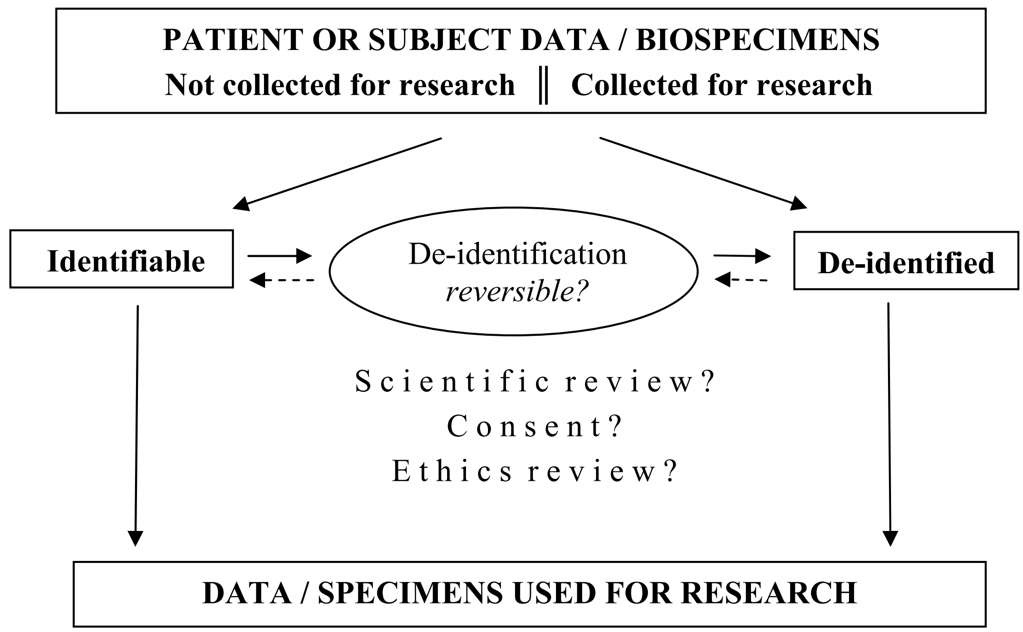
NHGRI Medical Sequencing Program (MSP). A program in which contributing investigators will submit samples and phenotypic data, and NHGRI will perform sequencing, maintain all the data in a database, and manage release of the data. Initial intentions are to sequence intervals associated with Mendelian disorders, and to sequence large numbers of samples from studies of complex disorders in order to gauge the distribution and frequency of medically relevant genes. Detailed planning and pilot analyses are underway. (http://www.genome.gov/15014882)

[43] NIH, "Proposed policy for sharing of data obtained in NIH supported or conducted genome-wide association studies (GWAS)," 71 *Federal Register*, 51629-51631 (August 30, 2006).

The Cancer Genome Atlas (TCGA). A proposed project to chart the inherited and acquired mutations that relate to the onset, diagnosis, progression, and treatment of cancers, by genotyping biospecimens and examining the genomic data in light of clinical data on the patients. Piloting is starting. (http://cancergenome.nih.gov)

UK Biobank. A project that at the end of 2006 will start recruiting 500,000 people around the U.K. in the age range 40−69, conduct physical examinations, collect biospecimens and lifestyle data, link to NHS medical records, and follow the health trajectories of the participants for several decades. Consent will be very broad and will cover possible genotyping. (www.ukbiobank.org)

Flowchart 1. Data flow when directly from a collection to researchers



Flowchart 2. Data flow when via a research resource platform to researchers