Published in final edited form as:

J Med Ethics. 2012 May ; 38(5): 304–309. doi:10.1136/medethics-2011-100181.

How anonymous is ‘anonymous’? Some suggestions towards a coherent universal coding system for genetic samples

Harald Schmidt¹ and Shawneequa Callier²

¹Center for Health Incentives, Leonard Davis Institute of Health Economics, University of Pennsylvania, Philadelphia, Pennsylvania, USA

²Clinical Research and Leadership, George Washington University School of Medicine, Washington, DC, USA

Abstract

So-called ‘anonymous’ tissue samples are widely used in research. Because they lack externally identifying information, they are viewed as useful in reconciling conflicts between the control, privacy and confidentiality interests of those from whom the samples originated and the public (or commercial) interest in carrying out research, as reflected in ‘consent or anonymise’ policies. High level guidance documents suggest that withdrawal of consent and samples and the provision of feedback are impossible in the case of anonymous samples. In view of recent developments in science and consumer-driven genomics the authors argue that such statements are misleading and only muddle complex ethical questions about possible entitlements to control over samples. The authors therefore propose that terms such as ‘anonymised’, ‘anonymous’ or ‘non-identifiable’ be removed entirely from documents describing research samples, especially from those aimed at the public. This is necessary as a matter of conceptual clarity and because failure to do so may jeopardise public trust in the governance of large scale databases. As there is wide variation in the taxonomy for tissue samples and no uniform national or international standards, the authors propose that a numeral-based universal coding system be implemented that focuses on specifying incremental levels of identifiability, rather than use terms that imply that the reidentification of research samples and associated actions are categorically impossible.

INTRODUCTION

Anonymity comes in many guises. It offers a range of advantages to, for example, witnesses in forensic proceedings, ‘whistle-blowers’ in professional settings, academic peer reviewers and financial donors. Implied in all these uses of anonymisation is a common sense meaning which is, roughly, that whatever was at some point directly linked to a person (an action, a statement, a testimony, an assessment or a sum of money) is unlinked by some process that makes it practically impossible to reidentify that person. For research purposes, using ‘anonymous’ or ‘anonymised’ biological samples is widely perceived as an appropriate way of reconciling conflicts between the control, privacy and confidentiality interests of those from whom the samples originated and the public (or commercial) interest in carrying out research.¹

Correspondence to Harald Schmidt, Center for Health Incentives and Behavioral Economics, Leonard Davis Institute of Health Economics, University of Pennsylvania, 1129 Blockley Hall, 423 Guardian Drive, Philadelphia, PA, US; schmidt@mail.med.upenn.edu.

Competing interests None.

Provenance and peer review Not commissioned; externally peer reviewed.

Misleading communications could jeopardise public trust in large-scale population databases that are of increasing importance in the advancement of genomic science and could hinder international research collaboration. We argue that harmonising the current taxonomy regarding the identifiability of research samples is necessary and possible in principle. We begin with a brief review of the characterisation of deidentified and anonymous samples in key policy documents, and discuss different ways in which linkage between samples and individuals from whom they originated can be established. We conclude that the concept of anonymous samples is inaccurate and potentially misleading and make a proposal for an alternative numeral system, in which degrees of identifiability are specified.

ANONYMOUS TISSUE SAMPLES, THE NATIONAL AND INTERNATIONAL POLICY CONTEXT

Regulatory consent policies typically only govern identifiable samples.² For example, in the USA and the UK, consent is usually required for the use of research samples that can be linked to the person from whom they came.¹ Where biological samples undergo a process of deidentification or anonymisation, however, individual consent is typically not required by law.^{3–5} This is reflected in the US Federal Policy for the Protection of Human Subjects which excludes specimens that are not ‘individually identifiable’ from its protective reach.⁶⁷ According to guidance issued by the Office for Human Research Protections (OHRP), samples are not identifiable ‘when they cannot be linked to specific individuals by the investigator(s) either directly or through coding systems.’⁸ Under the Health Insurance Portability and Accountability Act, ‘protected health information’ that is used by covered entities (eg, providers, health plans and healthcare clearing houses) is deidentified when all 18 of the enumerated identifiers are removed from the sample (including one’s name, birth date, social security number and medical record number); when protected health information is no longer reasonably considered to be ‘individually identifiable health information’; or when an expert in statistical or scientific methodologies believes that the risk of reidentification is small.⁹¹⁰ Equally, under the UK’s Human Tissue Act 2004, tissue from the living may be stored and/or used without consent for health-related research when such projects have been ethically approved and ‘the tissue is anonymised such that the researcher is not in possession of information identifying the person from whose body the material has come and is not likely to come into possession of it.’¹¹ Two related problems have been identified in the literature. First, the identifiability of a sample can shift as the sample changes hands, leading to intentional or inadvertent reidentification and making the current consent policies inadequate to address the reality of potential reidentification.¹⁰¹² Second, the legal characterisation of deidentified and anonymised samples is of a technical, rather than a conceptual nature. It does not suggest that linkage is impossible in principle, but that under some circumstances, professionals using a sample are not able to trace it to a particular individual. Yet, important guidelines on biological samples fail to consistently describe and provide adequate clarity regarding these nuances.^{13–14} Instead, as boxes 1 and 2 illustrate, there is considerable variation regarding the terms used to denote deidentified, anonymous or anonymised samples; the extent to which linkage is described as possible or impossible; and provisions on feedback or the ability to withdraw consent.

Clearly, the rather categorical statement by the International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use (ICH) declaring that in the case of anonymous samples there is no potential to trace them back to individuals and that it is not possible to enable sample withdrawal or feedback is insufficiently nuanced. The Human Genome Organization equally appears to envisage that there is a class of samples that are impossible to trace to a person, while the Medical Research Council takes a slightly more cautious stance by implying that there may be ways of linking samples, but that it is acceptable to assign the ‘anonymous’ status where the

information contained could not reasonably be used by anyone to establish the link. The Medical Research Council's distinction between linkage being established by 'the research team' as opposed to 'anyone' is not found explicitly in the other documents, which are concerned primarily with linkage options for the party holding the anonymous sample. Omitting this wider perspective in a global research community is problematic.

Provided that the biological or genetic material is of suitable quality and quantity, samples that contain DNA always, by their very nature, retain a link to the person from whom they came: 'DNA is itself uniquely identifiable.'²¹ Reidentification may require a great deal of effort, time and expense, as illustrated by the processes employed to identify victims of a natural disaster or terrorism (such as the Tsunami in the Indian Ocean in 2004 or the attacks on the World Trade Center in 2001). But it is wrong to suggest that such reidentification is impossible. As a number of recent scientific developments have demonstrated, it is increasingly easier to establish such matches. This trend looks set to continue, and the growing interconnectivity of increasingly larger and numerous genetic and non-genetic databases are of particular relevance, as are different ways in which people engaging in consumer-driven genomics might themselves initiate matching, as we review below.

ANONYMOUS TISSUE SAMPLES IN THE CONTEXT OF INCREASINGLY ACCESSIBLE AND INTERCONNECTED RESEARCH

Databases of biological or genetic samples are playing a growing role in healthcare research, and datasets are increasingly distributed and analysed through linked computer networks.^{22,23} As noted above, one of the motivations behind anonymising samples in the first place is to protect identity and privacy. Somewhat paradoxically though, the more that anonymous samples are used, the more the core notion of deidentification or anonymisation becomes questionable. For the clearer the phenotypic functions of particular genes become, the easier it becomes to make inferences about individual people whose DNA is included in pools of samples. At the same time, completely deidentified samples are less useful for research purposes.

The National Institutes of Health and the Wellcome Trust recently saw themselves forced to remove anonymous partial sequences of pooled DNA from their online datasets as a result of a study demonstrating that it is possible to identify DNA contributors to complex DNA mixtures even when their DNA makes up <0.1% of a mixture.^{24,25} It is plausible to assume that individuals who gave an anonymous sample for such research purposes would not expect this to be the case. However, since 'every individual shifts a genetic pool subtly in certain directions. studying enough single nucleotide polymorphisms (SNPs) unveils the pattern of those shifts'²⁴ which means that even in large scale research projects such as genome-wide associations studies there is 'a clear path for identifying whether specific individuals are within a study based on summary-level statistics.'²⁴ The National Institutes of Health have developed a new policy that requires researchers to apply for access to the aggregate data and to agree to protect the confidentiality of the data in the same way that it would for individual-level study, which provides improved confidentiality protections for the subjects.²⁶ Research has also demonstrated that it is possible to reidentify seemingly anonymous DNA sequences by linking them with other publicly available qualifiers such as gender, age or zip code and then matching the linked DNA with records containing further identifying information, such as census records. As Malin and Sweeny state, '87% (216 million of 248 million) of the population in the United States had reported characteristics that likely made them unique based only on [5 digit ZIP, gender, date of birth]'.²⁷ When the researchers matched the demographic data they obtained from hospital discharge records with census data, they 'reidentified' individuals who had a unique combination of demographic information.

Following this method, Malin and Sweeny²⁷ linked hospital discharge data with DNA sequences stored by various hospitals in Illinois between 1990 and 1997. Relying on a computer algorithm, they were able to search accessible DNA databases to discover the different hospitals that recorded one particular person's unique DNA sequence. Once they identified the hospitals where each person visited, they linked the patients' DNA sequences (which revealed disease information) with diagnostic codes and by extension demographic information stored in a hospital discharge database during the same time period. In the next step, zip code, gender and date of birth specific to each DNA entry were matched with named individuals found in a population register. As they explain, 'This work uses the mere existence of the DNA entry in multiple data sets to draw inferences about where the person has been. The person's visit pattern is then linked to other information to explicitly identify the person.'²⁷ The approach enabled the identification of 33%–100% of individuals who had specific genetic diseases, with the success rate having an inverse relationship to the number of patients with a particular disease.²⁷ In a separate, more confined, study it was possible to identify 98%–100% of individuals.²⁸

Linkage can also be established from outside of the research community, calling into question the core concept of an anonymous sample, in terms of the understanding in everyday language and in high level guidance, as described above. As Malin and Sweeny note, '44 of the 50 [US] states (or 88%) have legislative mandates to gather hospital-level data on each patient visit'²⁷ that in many cases have been distributed or sold to industry or made publicly available. Such linkage makes research participants vulnerable to third parties interested in retrieving identifiable information about genetic groups within a particular non-genetic database or demographic.

ANONYMOUS TISSUE SAMPLES IN THE CONTEXT OF CONSUMER DRIVEN GENOMICS

Another way in which seemingly anonymous samples can become reidentified is through action by those from whom the samples initially originated. Suppose that news coverage revealed that a study using anonymous samples had shown that people with a certain genetic marker were highly susceptible to a particular harmful condition. Because of the ICH terminology it might be argued that feedback could not be provided due to the samples' anonymous status. This may be true if we only draw on the information available to the research team. But clearly, on a slightly wider perspective this becomes questionable. In principle, individuals would merely need to provide another sample to allow for a rapid matching of their DNA sequences. Looking ahead, generating a new DNA profile might not even be necessary, as individuals might bring along some form of a digitalised 'personal genome'²⁹ comprising a full or partial representation of their genetic profile. Apart from services such as 23 and Me or Navigenics that offer different types of genome analysis, there are efforts underway towards making available at reasonable cost a full personal genome, sometimes referred to as the '\$1000 genome'.²⁹ It is presently not clear whether such services will include comprehensive disease susceptibility information or just a raw, un-interpreted DNA code. Factors such as the business case behind personal genome services and the state of scientific advancement are all likely to play a role.³⁰ However, even in the most conservative scenario, where a personal genome would simply include one's full, raw DNA code, it would be clear that a rapid 'matching exercise' against the DNA sequences of anonymous samples used in a research study would be highly feasible.

Recent empirical research supports the view that the public's opinions are not necessarily aligned with the assumptions behind governance arrangement, notably the basic assumption behind the ubiquitous 'consent or anonymise' policy.³¹ Further to earlier research,³² one might expect that patients' interest in knowing whether their samples were used in research

was strong in the case of identifiable samples, but far less strong, if not neutral, in the case of anonymous ones. But a study by Hull *et al*³¹ showed that while 81% wanted to know about research on identifiable samples, a somewhat unexpected 72% also wanted to know about research on their anonymous samples. Regarding control over anonymous samples, 56% thought it would be appropriate for researchers to require permission to carry out research, whereas 43% were content with notification. Hull *et al* noted that the reasons for wishing to be informed about research on anonymous samples were largely curiosity-driven (37%), with rights-claims also playing a role (19%). Meanwhile, 14% thought they should be able to prevent their sample from being used for research where the objective seemed frivolous or morally objectionable to them, and 17% wanted to know about the research uses of their sample to find out whether the results might benefit them personally, assuming that feedback could be provided readily.³¹

To be very clear: the mere fact that some percentage of patients desire feedback and/or would like to control the use of anonymous samples does not, by itself, mean that policy should be revised to enable this. For example, legitimate reasons for why it would be desirable to prevent withdrawal could include ensuring a suitably high number of specimens for effective public health disease surveillance. Policies that enable feedback or withdrawal or prevent certain research uses could also be costly or inefficient in implementation. With regard to sharing results, health risk information may not always be sufficiently robust or may require interpretation and/or counselling that physicians and researchers feel unable to provide adequately. Yet, providing feedback may be ethically desirable when critical and meaningful health information could be provided.

We have no intention of settling here any aspect of the highly complex question of patients' entitlement to control the research use of anonymous samples, of questions around the permissibility of feedback or the overall appropriateness of the consent or anonymise policy. However, what we wish to make clear is that it is wrong to muddle these ethical questions with scientific, conceptual or empirical claims about the status of such samples in terms of the possibility of establishing a link to the person from whom they originated.

CONCLUSION AND STEPS TOWARDS A POSSIBLE SOLUTION

As Gibbons observed, how we define 'biological materials, DNA or other genetic samples, and various categories of data' and how 'each concept is defined in the law (or not, as the case may be) can exert a major influence over its governance.'³³ In this context, terminology such as that advanced by the ICH is misleading and ought to be clarified. It is equally, if not more, important to ensure adequate language in patient and public information leaflets. For example, the UK Patient Information Advisory Group which explains anonymisation as 'turning information that identifies you into anonymous information through the removal of identifiers' is misleading.³⁴

So what should be done? As box 2 shows, Lowrance and Collins¹ and Knoppers and Saginur¹³ proposed to reduce the number of categories, suggesting that it might also be appropriate to specify fewer than five steps. But even if we follow this proposal, significant ambiguity and semantic confusion remains. The best way of avoiding this situation would be to remove terms such as 'deidentified,' 'anonymised' and 'anonymous' from policy documents at all levels, especially those aimed at patients or research participants. The same goes for 'non-identifiable' samples (box 2) or samples where links to individuals are 'irreversibly deidentified.' Instead, it would seem possible to adopt a numerical instead of a verbal classification system to signify different degrees of identifiability.

We therefore suggest that major research funders, such as, the US National Institutes of Health, and the Wellcome Trust and Medical Research Council in the UK in consultation with the international research community develop a single universal coding system to replace the unhelpful current variation of terms by focusing on specifying *incremental levels of identifiability*. For example, numerals I–V could cover the range of different degrees of possibility of linkage as represented in the categories found in boxes 1 and 2. Numeral I would refer to what is commonly defined as ‘identified’, and might include samples like those described in the ‘Open Consent’ model, where contributors can expect their samples and data to be published and shared in an identifiable form.³⁵ Numeral V would be assigned in those cases where it is most difficult to establish linkage, that is, all those currently labelled as ‘anonymous’. This category would therefore apply, for example, to samples that are used once, then destroyed, and only described in publications in the aggregate. Further, information to be disclosed during the consent process and security and data control requirements could be attached to the different stages. This approach would go some way towards avoiding misleading conceptions among, in particular, members of the public.

We realise, of course, that transitioning from the current system to a numerical-based one is no trivial matter, faces practical, structural and also political challenges, and cannot be implemented by simply replacing words with numerals. Yet, we are encouraged by the recent US efforts of OHRP to revise the decades old consent or anonymise policy. In July 2011, OHRP published an Advance Notice of Proposed Rulemaking that acknowledges the limitations of the Common Rule, including the inadequacy of the deidentification provisions.³⁶ Responding to concerns raised here and elsewhere—that public preferences are unaligned with the Common Rule³⁶ and that technological advancements are defying traditional conceptions of identifiability³⁶—OHRP proposes to terminate the consent or anonymise policy for research samples (but keep it in place for samples first collected for non-research purposes) and require open-ended, blanket consent for future uses of identifiable samples.³⁶ Although an Advance Notice of Proposed Rulemaking is only the first administrative step in the rule making process, the changes described above are evidence that OHRP recognises the limitations of anonymisation and is set to modify the rule with the preferences of sample contributors and researchers in mind.

We submit, however, that the proposed modifications only partially address the areas of concern described in our analysis. Presumably, and laudably, there would be greater awareness among sample contributors that their research samples and data would be reused in future research projects. Additionally, the legal standards for confidentiality would be greater and more consistent than they are now as the new rule would explicitly prohibit reidentification even in cases when it is technically possible. Finally, the return of results would be possible in the case of samples used for further research purposes when that research has been approved by an Institutional Review Board.³⁶

OHRP has not insisted, however, that all researchers and regulators cease using the term deidentified or anonymised to describe samples. Using these terms may still cause confusion as researchers who first collect samples for non-research purposes decide subjectively whether consent is required to repurpose the sample. Further, OHRP emphasises consent for inter-institutional uses but places much less emphasis on discussions about consent related to sample and data sharing among institutions, states and countries. Finally, there is little discussion about long term uses and repeated reuses of samples.

It also needs to be emphasised that these reform efforts are made in the USA only. We are not aware of similar initiatives in the UK or at the global level; clearly, however, the same issues arise outside of the USA too and in international research collaborations. It is to be hoped that a critical discussion will ensue elsewhere too. While we cannot address here

specific implementation issues, we suggest that a numerical system is worthy of consideration in these debates, as it is capable of communicating in a more appropriate way the potential risks and benefits associated with varying degrees of identifiability of samples.

We need to address three categorical objections to our proposal, however, before moving on. First, would a numeral system be intelligible to the public? Second, what if a numeric policy and more transparency about the incremental nature of identifiability lead more people to refuse research participation or donation of samples for research? Third, would the administrative burden of introducing and operating a new policy be proportionate to the potential gains?

While the proposal may be accepted as a more adequate reflection of the scientific scope and limitations of identification, it may be disputed that, in fact, it achieves the second and related goal, which is to improve transparency for research participants. For example, a point scale (with however many intervals) may be perceived as arcane and confusing by the public. This valid and important concern cannot be decided from the armchair, but requires careful pilot testing of suitable alternatives. Such testing should evaluate the current system against different numerical scales (say, three, five or ten point scales), and could also compare the use of alternative, more neutral verbal descriptors that clarify the continuum of ease of identifiability (eg, ranging from ‘easy’ to ‘extremely difficult’). We are not aware of any such research to date, and hence the claim that a numeral system would not be intelligible as well as the claim that the current system is, in fact intelligible, and superior to alternatives, awaits empirical assessments.

Regarding the related questions of the possible administrative burden and the impact on research participation, it is equally impossible to evaluate this in the abstract. However, it seems clear to us that even if the burden were considerable, it would also need to be balanced against the burden imposed by the risk of staying with the current situation: the prospect of a large scale public backlash and loss of trust in genetic research—and possibly in the medical research community more widely—does not suggest that the scales would tilt in favour of simply keeping the status quo. Equally, while a decrease in the willingness to participate in research or donate samples would clearly be regrettable, the alternative of sustaining levels where they are currently by leaving the taxonomy untouched might only yield a short-lived benefit. Hence, while it is more than possible that there are costs associated with a proposal such as the one sketched out here, the current situation is not cost free either.

In view of recent scientific developments, many crucial aspects in the current policies are problematic, most notably alleged claims on the impossibility of feedback and withdrawal. Not responding to these problems may undermine the trust of the public in increasingly larger and interconnected databases. In the absence of a more adequate system it is important neither to overstate nor to understate the implications of using samples that are labelled ‘anonymous’ or ‘anonymised’ in terms of withdrawal of consent or samples or provision of feedback. It would help to recognise more fully that the various stages of anonymisation are merely incremental levels that make it increasingly difficult to identify samples, but never impossible. While the exact route to better terminology will require further analysis than what we propose here, it is crucial that we move towards a system that eliminates the potential of misunderstandings about the (im)possibility of reidentifying allegedly anonymous samples which arise under the current terminology.

Acknowledgments

The authors are most grateful to Caroline Rogers for her insightful comments and participation in detailed discussions about earlier drafts of the manuscript. The authors are also grateful for discussion with and comments

from Stephanie Dyke, Sara Hull, Varsha Jagadesham, Graeme Laurie, Kathy Liddell, Mike Parker, Jeremy Sugarman, Julia Trusler and Hugh Whittall who reviewed earlier versions of this paper. The usual caveats apply. Partial support for this essay came from the Postdoctoral Fellowship program at the Center for Genetic Research Ethics and Law (Callier) through the National Institutes of Health grant P50-HG003390 from the National Human Genome Research Institute.

REFERENCES

1. Lowrance WW, Collins FS. Identifiability in genomic research. *Science*. 2007; 317:600–602. [PubMed: 17673640]
2. The Academy of Medical Sciences. Personal Data for Public Good: Using Health Information Medical Research. 2006 <http://www.acmedsci.ac.uk/download.php?file=/images/project/Personal.pdf>.
3. General Medical Council. Medical research: the role and responsibilities of doctors. *Bull Med Ethics*. 2001:18–24.
4. Kohane IS, Mandl KD, Taylor PL, et al. Medicine. Reestablishing the researcher-patient compact. *Science*. 2007; 316:836–837. [PubMed: 17495156]
5. [accessed 12 Aug 2011] Secretary's Advisory Committee on Genetics, Health Society Realizing Potential Pharmacogenomics Opportunities Challenges. 2008. http://oba.od.nih.gov/oba/SACGHS/reports/SACGHS_PGx_report.pdf
6. United States, Code of Federal Regulations, Title 45, Section 46.101(b)(4).
7. United States, Code of Federal Regulations, Title 45, Section 46.102(f)(2).
8. Office for Human Research Protections. [accessed 12 Aug 2011] Guidance Res Involving Coded Private Inf or Biol Specimens. 2008. <http://www.hhs.gov/ohrp/humansubjects/guidance/cdebiol04.htm>
9. United States, Code of Federal Regulations, Title 45, Section 164.514.
10. Rothstein MA. Is Deidentification Sufficient to Protect Health Policy in Research? *Am J Bioethics*. 2010; 10:3–11.
11. Human Tissue Authority. [accessed 12 Aug 2011] Code of Practice—Consent. 2006. http://www.hta.gov.uk/_db/_documents/2006-07-04_Aproved_by_Parliament_-_Code_of_Practice_1_-_Consent.pdf
12. Brothers KB, Clayton EW. Human Non-Subjects Research: Privacy and Compliance. *Am J Bioethics*. 2010; 10:15–17.
13. Knoppers BM, Saginur M. The Babel of genetic data terminology. *Nat Biotechnol*. 2005; 23:925–927. [PubMed: 16082354]
14. Medical Research Council. Human Tissue Biological Samples for Use Research—Operational Ethical Guidelines. 2001 http://www.mrc.ac.uk/consumption/idcplg?IdcService=GET_FILE&dID=9051&dDocName=MRC002420&allowInterrupt=1.
15. European Medicines Agency. Position Paper on Terminology in Pharmacogenetics. 2002 <http://www.emea.europa.eu/pdfs/human/press/pp/307001en.pdf>.
16. European Medicines Agency. Understanding the Terminology Used in Pharmacogenetics. 2004 <http://www.emea.europa.eu/pdfs/human/pharmacogenetics/384204en.pdf>.
17. Food and Drug Administration. [accessed 12 Aug 2011] Guidance for Industry, E15 Definitions for Genomic Biomarkers, Pharmacogenomics, Pharmacogenetics, Genomic Data and Sample Coding Categories. <http://www.fda.gov/downloads/RegulatoryInformation/Guidances/ucm129296.pdf>
18. The International Conference on Harmonisation (ICH) of Technical Requirements for Registration of Pharmaceuticals for Human Use; Definitions for Genomic Biomarkers, Pharmacogenomics, Pharmacogenetics, Genomic Data and Sample Coding Categories E15; 2007. http://www.ich.org/fileadmin/Public_Web_Site/ICH_Products/Guidelines/Efficacy/E15/Step4/E15_Guideline.pdf
19. Human Genome Organization Ethics Committee (HUGO Ethics Committee). Statement on DNA Sampling: Control and Access. 2002 http://www.hugo-international.org/img/dna_1998.pdf.

20. European Commission, The Independent expert Group. Ethical, Legal and social aspects of genetic testing: research, Development and Clinical Applications; 2004.
http://ec.europa.eu/research/conferences/2004/genetic/pdf/report_en.pdf
21. McGuire AL. Identifiability of DNA data: the need for consistent federal policy. *Am J Bioeth.* 2008; 8:75–76. [PubMed: 19003718]
22. National e-Science Centre. Defining e-Science. <http://www.nesc.ac.uk>.
23. McGuire AL, Gibbs RA. Genetics. No longer de-identified. *Science.* 2006; 312:370–371. [PubMed: 16627725]
24. Couzin J. Genetic privacy. Whole-genome data not anonymous, challenging assumptions. *Science.* 2008; 321:1278. [PubMed: 18772401]
25. Homer N, Szelinger S, Redman M, et al. Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. *PLoS Genet.* 2008; 4:e1000167.
26. National Institutes of Health. [accessed 11 Aug 2011] Background Fact Sheet on GWAS Policy Update. 2008. http://grants.nih.gov/grants/gwas/background_fact_sheet_20080828.pdf
27. Malin B, Sweeney L. Re-identification of DNA through an automated linkage process. *Proc AMIA Symp.* 2001:423–427. [PubMed: 11825223]
28. Malin B, Sweeney L. Determining the identifiability of DNA database entries. *Proc AMIA Symp.* 2000:537–541. [PubMed: 11079941]
29. Church GM. Genomes for all. (Cover story). *Sci Am.* 2006; 294:47–54.
30. Greenbaum D, Du J, Gerstein M. Genomic anonymity: have we already lost it? *Am J Bioeth.* 2008; 8:71–74. [PubMed: 19003717]
31. Hull SC, Sharp RR, Botkin JR, et al. Patients' views on identifiability of samples and informed consent for genetic research. *Am J Bioeth.* 2008; 8:62–70. [PubMed: 19003716]
32. Wendler D. What research with stored samples teaches us about research with human subjects. *Bioethics.* 2002; 16:33. [PubMed: 12061383]
33. Gibbons SM. Are UK genetic databases governed adequately? A comparative legal analysis. *Leg Stud.* 2007; 27:312–342.
34. Patient Information Advisory Group. [accessed 11 Aug 2011] Your Health Records Safeguarding confidential information - role patient information Advisory Group. 2006.
<http://www.advisorybodies.doh.gov.uk/piag/HealthRecords.pdf>
35. Lunshof JE, Chadwick R, Vorhaus DB, et al. From genetic privacy to open consent. *Nat Rev Genet.* 2008; 9:406–411. [PubMed: 18379574]
36. Department of Health and Human Services. [accessed 21 Jan 2012] Advance Notice of proposed Rulemaking, Human Subjects Research Protections: Enhancing Protections for Research Subjects and Reducing Burden, Delay, and Ambiguity for Investigators, 76 Fed. Reg. 44512.
<http://www.gpo.gov/fdsys/pkg/FR-2011-07-26/pdf/2011-18792.pdf>

Box 1 Examples of standards for classifying genetic or biological samples as 'deidentified,' 'anonymised' or 'anonymous.'

European Medicines Agency (EMA), US Food and Drug Administration (FDA) and International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use (ICH)

The EMA¹⁵¹⁶ and the FDA¹⁷ have adopted the five-pronged definitions developed in a 2007 Guideline by the ICH.¹⁸ Further to *identified*, *single-coded* and *double-coded* samples and data, the guideline characterises as follows:

- *Anonymised samples and data*: Those that are initially coded or double-coded but are no longer traceable back to the individual through coding keys. The link between the subjects' identifiers and the unique code(s) is subsequently deleted, which prevents subject reidentification through the use of coding keys.
- *Anonymous samples and data* are never labelled with personal identifiers and a coding key is not generated. There is 'no potential to trace back genomic data and samples to individual subjects.' Furthermore, it is 'not possible to undertake actions such as sample withdrawal, or the return of individual results, even at the subject's request.'¹⁸

Human Genome Organization (HUGO)

The HUGO has no specific guidance document on the use of genetic samples. However, its *Statement on DNA Sampling: Control and Access* (1998) notes:

'The choices offered in the consent process should reflect the potential uses of the DNA sample and its information. It is important to indicate whether the sample and its information will: identify the person, code the identity, or anonymize the identity so that the person cannot be traced although some demographic and clinical data may be provided.'¹⁹

European Commission

An expert group for the European Commission proposed a five-tiered approach in a 2004 report on genetic testing.²⁰ In addition to *identified*, *single-coded* and *double-coded* samples, the following two were characterised:

- *Anonymised*: Double-coded samples or data where the key linking the first and second code is destroyed. Can also include previously identified samples or data where the personal identifier has been destroyed.
- *Anonymous*: Those that do not have any personal identifiers or link with individual identity.

Medical Research Council (MRC) (UK)

In the UK, the MRC's Human Tissue and Biological Samples for Use in Research notes:

- *Operational and ethical guidelines* (2001) differentiate between two types of anonymised samples:
- *Linked anonymised samples and data*: 'fully anonymous to [...] the research team but contain information or codes that would allow others (e.g. the clinical team who collected them or an independent body entrusted with safekeeping of the code) to link them back to identifiable individuals.'

- *Unlinked anonymised samples and data:* ‘contain no information that could reasonably be used by anyone to identify the individuals who donated them or to whom they relate.’¹⁴

Box 2 Variation of terminology to describe biological or genetic samples

In a discussion document prepared for a workshop on privacy, confidentiality and identifiability in genomic research, convened by the National Human Genome Research Institute, William Lowrance and Collins commented on the wide variety of existing terms and proposed the following concordance of identifiability terms under a three-tiered structure:¹

Identified or identifiable: personal; nominative

Key-coded: reversibly deidentified; linked anonymised; pseudonymised; pseudoanonymised; encrypted; coded

Non-identifiable: irreversibly deidentified; unlinked anonymised; unidentifiable; anonymous.

In a commentary entitled 'The Babel of genetic data terminology,' Barbara Knoppers and Saginur equally reviewed the extant terminology and proposed a simplified two-tiered system to encompass otherwise used concepts:¹³

Coded (single/double): identifiable; linked; linked anonymised; potentially identifiable; reidentifiable; pseudonymised; reversibly deidentified/anonymised; proportional anonymity; reasonable anonymity; traceable; unidentified/unidentified for research purposes.

Anonymised: absolute anonymity; unlinked anonymised; non-identifiable; deidentified; irretrievably unlinked; irreversibly deidentified; irreversibly unlinked; non-identified; permanent anonymisation; unidentifiable; unlinked.