Original Paper

Public Health Genomics

Public Health Genomics 2011;14:17–25
DOI: 10.1159/000294150

Received: June 27, 2009
Accepted after revision: October 5, 2009
Published online: March 29, 2010

Assessing the Privacy Risks of Data Sharing in Genomics

C. Heeney[a] N. Hawkins[b] J. de Vries[a] P. Boddington[a] J. Kaye[b]

[a]The Ethox Centre, Department of Public Health and Primary Care, and [b]HeLEX – Centre for Health, Law and Emerging Technologies, Department of Public Health and Primary Care, University of Oxford, Oxford, UK

Key Words
Access to information · Confidentiality · Genetic privacy · Genetic research/ethics · Genome/human · Genomics · Humans · Information dissemination · Informed consent · Internet

Abstract
The protection of identity of participants in medical research has traditionally been guaranteed by the maintenance of the confidentiality of health information through mechanisms such as only releasing data in an aggregated form or after identifying variables have been removed. This protection of privacy is regarded as a fundamental principle of research ethics, through which the support of research participants and the public is maintained. Whilst this traditional model was adopted for genetics and genomics research, and was generally considered broadly fit for purpose, we argue that this approach is increasingly untenable in genomics. Privacy risk assessments need to have regard to the whole data environment, not merely the quality of the dataset to be released in isolation. As sources of data proliferate, issues of privacy protection are increasingly problematic in relation to the release of genomic data. However, we conclude that, by paying careful attention to potential pitfalls, scientific funders and researchers can take an important part in attempts to safeguard the public and ensure the continuation of potentially important scientific research.

The protection of identity of participants in medical research has traditionally been guaranteed by the maintenance of the confidentiality of health information through mechanisms such as only releasing data in an aggregated form or after identifying variables have been removed. This protection of privacy is regarded as a fundamental principle of research ethics, through which the support of research participants and the public is maintained. Whilst this traditional model was adopted for genetics and genomics research, and was generally considered broadly fit for purpose, various authors in the past few years have begun to question the effectiveness of the privacy protections in the genomic research context [1–6]. In 2008, the Wellcome Trust and the NIH removed open web access to genomic datasets after it was demonstrated that individuals could be re-identified from aggregated data on genome-wide association studies and that therefore the removal of identifying variables or the publishing of aggregate data alone were insufficient to protect the privacy of research participants [7–9].

Currently, in most areas of medical research the focus of privacy risk assessments tends to be on making the data non-identifiable, without extensive consideration of the existence of other datasets and resources and how they might increase the likelihood of identification. We

C.H. and N.H. contributed equally to this paper.

KARGER

Fax +41 61 306 12 34
E-Mail karger@karger.ch
www.karger.com

Accessible online at:
www.karger.com/phg

Catherine Heeney
The Ethox Centre, Department of Public Health and Primary Care
University of Oxford
Old Road Campus, Oxford OX3 7LF (UK)
Tel. +44 1865 287 887, Fax +44 1865 287 884, E-Mail catherine.heeney@ethox.ox.ac.uk

argue that this approach is increasingly untenable in genomics and draw on some important concepts from the field of statistics. Organisations such as National Statistical Institutes (NSIs) have long adopted proactive techniques to identify challenges to the privacy of individuals. We believe that insights from this approach to disclosure control hold important lessons for the field of genomics. Privacy risk assessments need to have regard to the whole data environment, not merely the quality of the dataset to be released in isolation. A key factor in this approach is the assumption of the existence of a data intruder – an individual intent on attacking the data for the purposes of reidentifying data subjects [10]. Statistical datasets released by NSIs are submitted to rigorous tests to examine the possibilities open to a data intruder for successfully identifying participants. We believe that this approach could be usefully employed when conducting a privacy risk assessment in the field of genomics. In this paper we will discuss the application of the data environment and data intruder concepts to genomics research and reflect on implications for governance in this field [5, 11].

The Data Environment

The assessment of disclosure risks and the potential to identify individuals should be based on an understanding of the type of information that is available in the data environment – the parameters of available data. (The concept of the data environment is employed in a similar way by the Data Environment Analysis Service based at Manchester University). Consideration of the data environment allows a systematic approach to the identification of sources beyond a given dataset and how they might enable reidentification of individuals in the original dataset. The data environment is therefore a useful concept if one is trying to move away from the practice of considering the risks of reidentification relating to the release of one dataset in isolation [10]. The rationale underlying this work in the field of disclosure control is that taking only the quality of released data into account is not enough to protect the privacy of data subjects, as it fails to acknowledge the usefulness of other available datasets for the purposes of reidentification of an individual. This resonates with recent challenges to privacy in genomic research, which relate not only to the possibility of identifying participants directly from available data but also to the identification of participants indirectly, by using data available from other sources. This challenge is not new, nor is it specific to genomics; the challenges to anonymity posed by potential inference from other datasets have long been debated with regard to the management and reuse of large statistical datasets, such as those released by NSIs [1, 2]. However, particular features of genomics, including the proliferation of high resolution genomic data and data sharing policies coupled with important recent developments in technology, are important factors to consider alongside the concept of the data environment.

Homer et al. used a process of comparing 3 datasets, 'the complex DNA mixture', 'a reference population' and the 'individual's genotype' to identify an individual in the data [7]. Here the data environment for the release of the anonymised datasets discussed in these papers included what was described by Homer et al. as the reference population. Genomic data is very high resolution and therefore is more likely than socio-economic data to provide a combination of traits that are individually unique. It is estimated that 75 – 100 single nucleotide polymorphisms or fewer than 20 microsatellite markers can unambiguously identify a single individual [12].

Genomic sequence data is being generated at an ever increasing rate. Research projects all over the world are generating genomic data by genotyping or sequencing the genome of their participants. This is happening not only in the research community but also in the private sector; sequence data is now being generated by a number of private companies who offer direct to consumer genetic testing and, in some cases, feed back raw sequence data as well as their interpretation of it [13]. Another source of data which has an impact on risks posed by the data environment for the release of genomic data is that produced by ancestor tracing companies or held in genealogical registries [14]. Many people may voluntarily submit their DNA to sequencing companies without being aware that these provide potential reference collections. Furthermore, where archived biological research samples are reused for genomic research, the participant may be unaware that their DNA is part of a collection being used for genomic research let alone the privacy implications of its release and reuse [5].

Genomic data is widely shared. Since the start of the Human Genome Project the importance of the sharing of sequence data for the advancement of science has been stressed by both funding bodies and influential scientists alike [15, 16]. On this basis, a number of projects have been established with the explicit aim of generating data that could be used by the scientific community as virtual reference libraries. Examples of projects that followed the open access model, which had proved so successful in the

Human Genome Project, are the HapMap project and more recently the 1000 Genomes Project.

Even though some restrictions on access to data within genomics remain, there is a movement towards releasing more data and even unlocking data sources which were previously available only in more restricted ways [17]. Initiatives such as UK Biobank will create a data resource, which will be open to researchers from public and private sectors and which will source data from the National Health Service under certain circumstances. Funders in the USA and the UK have invested in the creation of datasets to facilitate genome-wide association studies. These are for instance dbGaP, the Genetic Association Information Network (GAIN), and the European Genotype Archive. Projects such as the Wellcome Trust Case Control Consortium source data from existing projects such as the 1958 Birth Cohort and make these available on application. Some of the data, which until the publication of the Homer et al. paper were believed to be effectively anonymised, were publicly accessible on the web. More sensitive information or disclosive information were restricted and only available with the approval of a Data Access Committee. Now all data are subject to this level of control.

The data environment also includes other publicly available data sources relating, for example, to socio-economic data. Data drawn from publicly available identified datasets, such as voter registration or other forms of anonymised data such as census data, is also part of the data environment and can be used in combination with other data sources to reidentify individuals in specific data sets [1]. As well as census data, other data sources such as lifestyle databases created by supermarkets or data processing technologies like Geographical Information Systems (GIS), can help in isolating smaller groups of individuals within a dataset and facilitate disclosure of identities or attributes (traits). The work of disclosure control experts is constantly growing in sophistication to take account of the threats posed by new technologies and multiple overlapping data sources [18]. The internet plays a part here by increasing the number of available datasets, leading some disclosure control experts to strike a pessimistic note on the possibility of releasing 'safe' data: 'it is becoming increasingly difficult to produce anonymous and declassified information in today's globally networked society' [19].

The fact that data are available in different places and at different levels of aggregation and anonymisation provides less protection for privacy than one might think in the face of new techniques and technologies for comparing and analysing datasets. In this respect, advances in mathematical and statistical techniques, computational power and knowledge of genomics are all relevant. In a recent study, a combination of information from genealogical registries and haplotype analysis of the Y chromosome collected for the HapMap project allowed the prediction of the surnames of a number of individuals held in the HapMap dataset [14]. Nyholt et al. describe how advanced computational tools, together with increased knowledge of genomic structure, supported the inference of withheld information relating to James Watson's *ApoE* gene that had been withheld from published data [20]. Information technology, moreover, facilitates such comparisons, allowing individual data to be compared with statistical norms in order to determine how different a person is from a given population, a technique used extensively in the private sector [21, 22]. This comparison to a statistical norm creates further data relating to an individual thereby adding to the information available; it could, therefore, increase the possibility of disclosure of a person's identity. Again, the possibility of combining different relevant data sources coupled with certain types of expertise and data processing technologies facilitates the discovery of identities and characteristics of data subjects. Genomic data also provides information about family members [23] and other genetically related groups.

What is also crucial here is the vast increase in informational power unleashed by using the method of comparing and inferring from available data sources. For organisations with the computational capabilities to automate at least parts of this process, there are few barriers to the routine inference or discovery of information relating to individuals. Technologies such as data-mining do exactly this and have been recognised for a number of years as creating challenges to privacy [24]. When data-mining is used in combination with data on populations generated by genomic research, it has the capacity to enable 'the construction of new groups (based on arbitrary and non-obvious patterns and statistical correlations)' [2].

The combination of these particular factors of the data environment (the type and proliferation of data, and the statistical and technological advances) mean that traditional ways of protecting privacy are less effective in protecting the identities of research subjects. Traditional methods of statistical disclosure control often involve modifying data, through removing information by combining variables or by adding noise to the data. However, these techniques are often ineffective in a real world set-

ting where the combination of a number of data sets allows these traditional privacy protections to be circumvented [1].

The Data Intruder

The notion of a data intruder should be understood simply as someone with a motivation to investigate the attributes or identity of a data subject, and who uses available information for reidentification of individuals. Whilst it may seem that few would have the necessary skills, motivation or equipment to do this, we suggest that the data environment of genomics in fact includes many people who may be so motivated, and for various reasons. Their reasons might include further research, forensic purposes or use in marketing, insurance or employment decisions. Their motivations need not be sinister. Not only are there many who are driven by curiosity concerning genealogy or more widely, ancestry, but there are thousands with stronger motivations, including adoptees and donor conceived children [25, 26]. Many of the latter are, moreover, fiercely critical of current regulations, in various jurisdictions, that prevent them from discovering their genetic relatives. There is therefore a potential for some individuals to consider that they have rights to access genetic information which outweigh other ethical considerations such as confidentiality and privacy. Given the large numbers of interested and involved people, it would be foolish to consider that none of these would be motivated to stretch the boundaries of what constitutes a legitimate search for information.

Controlling Disclosure of Identity

Identification by a data intruder can arise in different ways from use of sources within the data environment. Direct identification of individuals in a single dataset of genomic information does not tend to occur because personal identifiers are removed in order to comply with data protection requirements in the relevant jurisdiction, which generally require that only anonymised data be released. However, in a data environment a dataset does not exist in isolation; datasets overlap and people and organisations have all sorts of knowledge, some of which can be difficult to predict. While direct identification of an individual from one dataset may be avoided by removing identifiers, indirect identification or identification via inference is still (re)identification and therefore a threat to privacy [2]. As a result, identifiability is not exclusively a quality of the data itself but depends also on the data environment and the motivations and resources available to a potential data intruder in any given situation.

The quality of the data is not unimportant and may have a bearing on how difficult it will be ultimately to reidentify a data subject. One example from the disclosure control field is provided by datasets which contain special uniques, i.e. individuals (or other statistical units, such as a family or a household) which are unique in a population. Statistical units may be easily identifiable because of the rareness of their characteristics or combinations thereof. The threats to anonymity posed by unique combinations of traits are well documented [27].

An example of a special unique would be a 24-year-old widowed female with an unusual genetic condition living in a small town, who would, in most cases, easily be identifiable in a dataset unless special steps were taken to avoid this. Again, information about the actual real world small town and its inhabitants would be an essential tool for connecting the data subject with the real world individual. However, the data environment for the release of this particular dataset includes any number of other information sources such as newspaper articles or even lifestyle databases assembled by supermarkets, depending on who the data intruder is assumed to be and which resources might be available to them. The possibility of identification of an individual in a given dataset containing anonymised or aggregated data seems less remote when one considers that no dataset exists in a vacuum and even a potential data intruder's own knowledge could allow reidentification.

The reidentification of individuals is not the only possible form of disclosure. Attribute disclosure is an associated type of disclosure which happens when an individual or group is known to be represented in a dataset which shows that everybody has a particular trait [28]. For example, a dataset could reveal that every house in a particular geographical area has 2 bathrooms. Therefore, if one had access to a person's address one would know from the dataset whether they had 2 bathrooms. This process is aided by the existence of other datasets which provide data on the same population. When 100% or none of the population represented in a dataset has a given trait, it can then be inferred with certainty that a person from that population does or does not have the trait.

However, in the real world it is more likely that a dataset shows that a large proportion of the population in question have the trait. Where for example 80% of a population has a particular trait, then an individual from this

population is very likely to have the trait. Other available datasets in combination with local knowledge and data analysis techniques such as data mining can further narrow the odds. Rather than having directly identified an individual in a dataset, a number of datasets are used to impute the characteristics about which knowledge is sought for that particular individual. The type of data used for this process could be genomic – Nyholt et al. were able to use a similar process to infer information about Watson's ApoE status. If a dataset provides the information that there is for example an 80% chance that an individual has a particular trait, it can significantly aid the process of inference. Although this might not be attribute disclosure in its pure form, it is useful information for a potential data intruder. Moreover, less than fully accurate predictions about individuals in a given population may nonetheless be used as a basis for making decisions. As with non-genomic data, concerns are raised about the use of this type of data by insurance companies, employers and others for genetic discrimination if individuals in disease study cohorts are identified [29].

Uses of the Results of Identification – What Are the Risks?

The perennial concern for those who take part in genomics research is that their genomic information may be used to discriminate against them by insurers or employers [29]. Many genomics studies concern health conditions, such as mental illnesses, to which society still attaches stigma. Although some jurisdictions have legislative protection against some forms of discrimination, not all jurisdictions provide such protection, which may, at any rate, be fairly limited in its scope. It is not a sufficient defence to claim that the information disclosed by most genomic data about risks for future health is not relevant on an individual basis. Moreover, researchers cannot assume that because genomic information does not disclose information about an individual today that it will not do so in the future. As research proceeds, it is reasonable to assume that more will be known about genetics and sequence information of an individual will be more informative.

Redlining is a process that uses profiling to exclude individuals from access to goods and services [30]. This process occupies a legal grey area in most jurisdictions, and it is, moreover, usually difficult to establish that it has happened [24]. Individuals and organisations who reuse data for redlining are data intruders who are not predominantly concerned about accuracy with regard to individuals but rather about making a well-informed guess. Redlining relies on the application of non-distributive profiles [31]. This means applying a profile based on the traits of some of a group to all members of the group, even though an individual member of a group may not have the undesirable trait. For example, an individual could be excluded from an insurance policy based on the fact that the majority of individuals living in the same geographical area had suffered from unusually poor health. The use of non-distributive profiles has had detrimental consequences in the form of discrimination for individuals where traits are strongly correlated to membership of certain racial or ethnic groups. Well-known examples concern African Americans who were stigmatised in relation to sickle cell disease in the 1970s [32]. Discrimination against those with sickle cell trait spread to carriers who were sometimes denied employment and life insurance. Genetic information may add its weight to the converging of disease stigma and racial discrimination [33].

As new discoveries in genomics are made public, there is further potential for discrimination of this kind. A more contemporary example concerns the Duffy antigen, which confers resistance to certain forms of malaria and is widely spread in African populations. A recent study also linked the Duffy antigen gene with heightened odds of acquiring HIV-1 [34]. Non-transparent allocation of individuals to groups based on known or inferred traits or some combination thereof can raise issues related to the ability to protect one's own interest and avoid discrimination, concerns traditionally associated with privacy [24].

Another potential use of genomic data is in forensics; the use of DNA evidence is an important tool in criminal investigation and can help to secure convictions. The most usual means of use is through matching of crime scene samples to profiles in a criminal DNA database, such as the UK Police National DNA database. These databases are usually compiled from samples taken from those who have been convicted of certain crimes and are subject to tight regulation [35]. Even so, the UK Police National DNA Database has been criticised on privacy and discrimination grounds. Other less tightly regulated collections of genomic data could also be used for reidentification for forensic purposes; law enforcement agencies have in the past matched DNA recovered from a crime scene to an individual's data in a biomedical dataset [36–38].

When DNA sequence information is freely available to all on the internet, it is a relatively easy matter for police

to search these biomedical datasets for matches for their crime scene samples. Once a match within a particular biomedical dataset is made, then it seems likely that the individual to whom the sample belongs would be identified, through either the release of the information by the database managers or following a court order. It is also worth noting that the police already use familial identification techniques, so sample matches to family members may be sufficient for identification [39]. Criminal profile matching does not necessarily identify only a single individual; it may identify a pool of suspects who are then all contacted to see whether they could be the person to be convicted. As more samples are searched, more false positives are likely to arise. When this is coupled with the fact that juries are inclined to misunderstand scientific evidence, particularly that relating to DNA, miscarriages of justice may arise [40].

Moreover, such uses may not be transparent; a researcher could release information to police or a court order could be made, and the means of identification need not become public. Although it may be argued that forensic use is justified in some circumstances, as the costs of DNA collection and sequencing drop this approach may be used with less serious crimes and in jurisdictions without robust protections for the rights of the accused. Whatever the rights and wrongs of such secondary uses, they certainly conflict with the original intention to create a database for genomic research purposes.

Some jurisdictions have regulatory protections which are intended to stop the unauthorised use of biomedical datasets, for example, for forensic use. In the USA, a certificate of confidentiality is meant to prevent this type of use of a research dataset. However, it is becoming increasingly clear that these protections may not actually work in practice [41]. The majority of countries lack even these protections.

Recommendations for Privacy Protections

Although the ability to share and access data is vital to the progress of scientific research, we should not forget the implications that a lack of protection for privacy could have for the lives of individual citizens. A clear recommendation is that it is no longer reasonable to assume that because data is anonymous in one data environment or at the present time, it will remain so in every data environment into the future. It is also not reasonable to assume that use of identification from genomic information will be innocuous.

Traditionally, legal frameworks, at least in liberal democracies, have sought to balance the privacy of data-subjects with the benefits of research by relying heavily on anonymisation and informed consent [4, 6]. However, the effectiveness of these twin tools for this purpose is compromised in the light of the discussion above. The current framework for protecting informational privacy assumes that the use of genomic datasets, or at least the resources to make use of them, would largely be restricted to the scientific research community; this is a naïve assumption. How the resources available to a data intruder will be used in each specific situation is difficult to predict, however, one thing that the studies above show is that this can happen and will happen. The current climate for data release is rich not only in datasets but also in knowledge, skills and motivations. In fact, a great many sources of information which are related to genetic information are widely available [11], as are the information processing technologies needed to make sense of the data [2]. Moreover, the form of potential privacy infringement, which relies on inference, lacks transparency precisely because it is indirect and carried out within specific data environments. As a result it is difficult both to detect and to guard against by considering only the quality of the data itself.

We are constantly warned to keep other personal information such as our names, phone numbers and bank account details as private as possible, to avoid identity theft. If privacy is breached in this way the situation can still sometimes be addressed, identity theft insurance may be purchased, and new accounts can be set up. Individuals cannot, however, change their DNA sequence: it is amongst the most personal of information. In the digital age, once genomic data is publicly released, it is virtually impossible to retrieve it or to make it private again, or even to know who has the information or to what use it is being put. This is especially so where some combination of inference and identifiable data is involved.

Restricting access to genomic data to legitimate academic researchers will go a long way towards reducing the privacy risks elaborated above. There could, for example, be a system which parallels that of the UK Data Archive which releases more sensitive and disclosive datasets only to licensed academics or researchers. Corporate entities will often be carrying out legitimate academic research in biomedical sciences, as in the social sciences. However, the profit motives and commercial imperatives of the private sector could also give rise to activities that cause particular privacy concerns. We recommend that it is therefore appropriate to consider care-

fully how commercial and potentially forensic use of this data is to be governed.

We conclude that privacy risk assessments carried out prior to the establishment of collections now need to take into account the resources available to potential data intruders in the data environment, rather than the quality of the data in an individual dataset in isolation. Genomic data collections will be long standing, and participants need to be protected long term, not only in the first years of their samples being held. The traditional focus of privacy protection in research, on consent and anonymisation, is incapable alone of addressing the concerns raised in this paper. Changes to the system of governance will help to address the concerns raised here, but a new system would need to take into account a broader context for privacy risks and not only the narrow confines of one project or the activities of the research community in isolation.

The recent Thomas and Walport report on data sharing [42] makes a number of recommendations regarding the changes to the law that should be considered in order to facilitate research within the UK. These include the establishment of safe havens for researchers and a system of approving and accrediting researchers. They recommend looking at the governance models that have been developed for statistical research as a basis for thinking about the use of medical information in research. While this is welcome, currently there are no policy requirements within medical research to undertake such broader assessments of privacy.

Importantly, it is clear that the promises of confidentiality made to participants in consent forms may need to be updated. It is both unrealistic and irresponsible to promise absolute confidentiality to participants in genomics research, where data is shared. This in turn raises the question of what to do about past promises in consent forms. Given that recontacting many thousands of participants may not be practicable, we recommend that robust assurances of realistic levels of protection for individuals and open public discussion, perhaps relating to the appropriate reuse of genomic data and related data, would be welcome.

The initial reaction of funders to the privacy challenge to which they were alerted by Homer and colleagues was to remove full open online access [9]. This was an appropriate initial response. Now is the right time to reconsider the implications of the release of genomic data for privacy. One response might be to accept that participants in genomic research cannot maintain their privacy; however, privacy risks can be properly assessed and steps taken to guard against them. In reality, the privacy risks associated with genomics research are part of a broader picture. Genomics researchers can only do so much to protect privacy. Ultimately, many of the concerns discussed above relate to matters such as insurance and employment; these concerns are really only satisfactorily addressed by appropriate legislative and governance responses, at a higher policy level, and by ensuring fair access to employment and healthcare.

The costs of a failure to reflect on these issues could be a fundamental loss of public trust, which is likely to affect willingness to take part in research. Again lessons can be learnt from the experience of the production and use of official statistics. For instance, a loss of trust was thought to be related to a fall in participation in the 2001 UK Census. This fall off was, moreover, not related to direct identification of individuals in datasets but rather a growing sense of unease among members of the public about the circulation and use of their data [43, 44].

Participants are an essential partner in genomic research, and without samples the research will not proceed. To have large numbers of individuals requesting that their samples be withdrawn could produce an expensive, time consuming problem for researchers. As sources of data proliferate, issues of privacy protection are increasingly problematic in relation to the release of genomic data. By paying careful attention to potential pitfalls, scientific funders and researchers can take an important part in attempts to safeguard the public and ensure the continuation of potentially important scientific research.

Glossary of Terms

Anomymisation is a process which involves removing identifiers from data. This can be done in a number of different ways, such as by eliminating variables and often by the removal of direct identifiers from the data. Anonymisation aims to minimize the risk of identity disclosure.

Attribute disclosure is attribution independent of identification. This form of disclosure is of concern to NSIs involved in tabular data release and arises from the presence of empty cells either in a released table or linkable set of tables after any subtraction has taken place. The presence of a single zero within a table means that a data intruder may infer from mere knowledge that an individual is represented in the table that the individual does not possess the combination of attributes within the cell containing the zero.

Direct identification occurs where a person's identity can be determined on the basis of the information about variables provided in a dataset.

Indirect identification/disclosure by inference occurs where a person can be identified not directly from one dataset but by a process of inference aided by other data sources which overlap with or explain aspects of the first dataset.

A *non-distributive profile* is an information profile built up from statistical or aggregated data about a population. Each person in the population will have a probability of having a constellation of traits. The profile of the average person this produces is not distributed in the same way for all members but may be applied as if it is.

Profiling involves collating information often derived from a number of resources to build profiles on individuals which are models that predict behaviour. These profiles may be used by marketers for target advertising. Companies may link profiles to individuals' identities.

Redlining is the practice of limiting financial or retail services to individuals who are part of a particular group. This is generally done on the basis of residents having a lower social status or income.

Statistical disclosure control techniques is the set of methods used to reduce the risk of disclosing information on individuals or organizations, usually by perturbing the data, or not releasing more disclosive data which could allow the identity of a statistical unit to be determined.

Acknowledgements

C.H. works on IBDchip, a project funded by the European Commission through grant number LSHB-CT-2006-037319. N.H., J.d.V. and J.K. are funded by the Wellcome Trust, under grant codes WT 077869/Z/05/Z, WT 083326/Z/07/Z, WT 081407/Z/06/Z and WT 076070/Z/04/Z. P.B. is funded by EU FP6, Procardis Project number LSHM-CT-2007-037273. This study was carried out as part of the work of the P3G Core on Data Sharing. P3G is funded by Genome Canada and Genome Quebec.

References

1 Malin B, Sweeney L: How (not) to protect genomic data privacy in a distributed network: using trail re-identification to evaluate and design anonymity protection systems. J Biomed Inform 2004;37:179–192.
2 Tavani HT: Genomic research and data-mining technology implications for personal privacy and informed consent. Ethics Inf Technol 2004;6:15–28.
3 Foster MW, Sharp RR: Ethical issues in medical-sequencing research: implications of genotype-phenotype studies for individuals and populations. Hum Mol Genet 2006; 15(spec No 1):R45–R49.
4 Lowrance WW, Collins FS: Identifiability in genomic research. Science 2007;317:600–602.
5 Lunshof JE, Chadwick R, Vorhaus DB, Church GM: From genetic privacy to open consent. Nat Rev Genet 2008;9:406–411.
6 Taylor P: When consent gets in the way. Nature 2008;456:32–33.
7 Homer N, Szelinger S, Redman M, Duggan D, Tembe W, Muehling J, Pearson JV, Stephan DA, Nelson SF, Craig DW: Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. PLoS Genet 2008;4:e1000167.
8 Couzin J: Genetic privacy. Whole-genome data not anonymous, challenging assumptions. Science 2008;321:1278.
9 Anonymous: DNA databases shut after identities compromised. Nature 2008;455:13.
10 Elliot MJ, Dale A: Scenarios of attack: the data intruder's perspective on statistical disclosure risk. Netherlands Official Statistics 1999;14:6–10.
11 Greenbaum D, Du J, Gerstein M: Genomic anonymity: have we already lost it? Am J Bioeth 2008;8:71–74.
12 Lin Z, Owen AB, Altman RB: Genetics. Genomic research and human subject privacy. Science 2004;305:183.
13 Kaye J: The regulation of direct-to-consumer genetic tests. Hum Mol Genet 2008;17:R180–R183.
14 Gitschier J: Inferential genotyping of Y chromosomes in Latter-Day Saints founders and comparison to Utah samples in the HapMap project. Am J Hum Genet 2009;84:251–258.
15 Sharing Data from Large-scale Biological Research Projects: A system of tripartite responsibility. Available at http://www.genome.gov/Pages/Research/WellcomeReport0303.pdf.
16 Le Monde Diplomatic: Nobel Prize for discoveries in genetics – heritage of humanity. Available at http://mondediplo.com/2002/12/15genome.
17 Kaye J, Heeney C, Hawkins N, de Vries J, Boddington P: Data sharing in genomics: reshaping scientific practice. Nat Rev Genet 2009;10:331–335.
18 Torra V, Domingo-Ferrer J, Torres A: Data mining methods for linking data coming from several sources. 3rd Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality. Luxembourg, 2003.
19 Sweeney L: Information explosion, confidentiality; in Doyle P, Lane JL, Theeuwes JJM, Zayatz L (eds): Confidentiality, Disclosure, and Data Access: Theory and Practical Application for Statistical Agencies. Amsterdam, Elsevier, 2001.
20 Nyholt DR, Yu CE, Visscher PM: On Jim Watson's APOE status: genetic information is hard to hide. Eur J Hum Genet 2009;17: 147–149.
21 Heeney C: Privacy and the identity gap in socio-technical system; in Whitworth B, de Moor A (eds): Handbook of Research on Socio-Technical Design and Social Networking Systems. IGB-Global, 2009, pp 110–122.
22 Nissenbaum H: Privacy as contextual integrity. Wash Law Rev 2004;79:119–157.
23 Cassa CA, Schmidt B, Kohane IS, Mandl KD: My sister's keeper? Genomic research and the identifiability of siblings. BMC Med Genomics 2008;1:32.
24 Vedder A: KDD, privacy, individuality, and fairness; in Spinello RA, Tavani HT (eds): Readings in Cyberethics. Boston, Jones and Bartlett Publishers, 2001, pp 404–412.
25 Bastard Nation – Search: Research, roots and reconnection. Available at http://www.bastards.org/search/.

26 The donor sibling registry. Available at http://donorsiblingregistry.com/.
27 Elliot M, Skinner CJ, Dale A: Special uniques, random uniques and sticky populations: some counterintuitive effects of geographical detail on disclosure risk. Research in Official Statistics 1998;1:53–67.
28 Organisation for Economic Co-operation and Development, SourceOECD (Online service): OECD glossary of statistical terms. Paris, OECD, 2008.
29 Ashcroft R: Should genetic information be disclosed to insurers? No. BMJ 2007;334: 1197.
30 6 P, Jupp B: Divided by Information? The 'Digital Divide' That Really Matters and the Implications of the New Meritocracy. London, Demos, 2001.
31 Vedder A: Medical data, new information technologies and the need for normative principles other than privacy rules; in Freeman MD, Lewis AD (eds): Law and Medicine: Current Legal Issues. Oxford, Oxford University Press, 2000, pp 441–459.
32 Bonham VL, Warshauer-Baker E, Collins FS: Race and ethnicity in the genome era: the complexity of the constructs. Am Psychol 2005;60:9–15.
33 Wailoo K: Stigma, race, and disease in 20th century America. Lancet 2006;367:531–533.
34 He W, Neil S, Kulkarni H, Wright E, Agan BK, Marconi VC, Dolan MJ, Weiss RA, Ahuja SK: Duffy antigen receptor for chemokines mediates trans-infection of HIV-1 from red blood cells to target cells and affects HIV-AIDS susceptibility. Cell Host Microbe 2008;4:52–62.
35 S v United Kingdom (30562/04) Marper v United Kingdom (30566/04) (2008) 25 BHRC 557. European Court of Human Rights.
36 Kaye J: Police access to DNA samples and information. Genomics Soc Policy 2006;2:16–27.
37 Hansson SO: The ethics of biobanks. Camb Q Healthc Ethics 2004;13:319–326.
38 Campbell AV: The ethical challenges of genetic databases: safeguarding altruism and trust. Kings Law J 2007;18:227–245.
39 R v Lloyd (James) [2007] EWCA Crim 1388. Court of Appeal (Criminal Division), UK.
40 Staley K: The Police National DNA Database: Balancing Crime Detection, Human Rights and Privacy. Genewatch UK, Buxton, 2005.
41 Beskow LM, Dame L, Costello EJ: Research ethics. Certificates of confidentiality and compelled disclosure of data. Science 2008; 322:1054–1055.
42 Thomas R, Walport M: Data Sharing Review Report. 2008.
43 Cohen N: Our missing million. The Observer, November 9, 2009.
44 Gerber ER: The privacy context of survey response: an ethnographic account; in Doyle P, Lane JL, Theeuwes JJM, Zayatz L (eds): Confidentiality, Disclosure, and Data Access: Theory and Practical Application for Statistical Agencies. Amsterdam, Elsevier, 2001.