EJHG Open

European Journal of Human Genetics (2016) 24, 1553–1558

www.nature.com/ejhg

ARTICLE

The risk of re-identification versus the need to identify individuals in rare disease research

Mats G Hansson*[,1], Hanns Lochmüller[2], Olaf Riess[3], Franz Schaefer[4], Michael Orth[5], Yaffa Rubinstein[6], Caron Molster[7], Hugh Dawkins[7,8,9,10], Domenica Taruscio[11], Manuel Posada[12] and Simon Woods[13]

There is a growing concern in the ethics literature and among policy makers that de-identification or coding of personal data and biospecimens is not sufficient for protecting research subjects from privacy invasions and possible breaches of confidentiality due to the possibility of unauthorized re-identification. At the same time, there is a need in medical science to be able to identify individual patients. In particular for rare disease research there is a special and well-documented need for research collaboration so that data and biosamples from multiple independent studies can be shared across borders. In this article, we identify the needs and arguments related to de-identification and re-identification of patients and research subjects and suggest how the different needs may be balanced within a framework of using unique encrypted identifiers.
European Journal of Human Genetics (2016) **24,** 1553–1558; doi:10.1038/ejhg.2016.52; published online 25 May 2016

INTRODUCTION

In an era of genomics, rapid technological advances, and globalization there are new challenges for protecting the identity and the privacy of the patient and balancing privacy interests against the needs of particular patient groups or societal interests; specifically for patients with rare diseases. These developments have resulted in a rapid increase in the volume of data collection and secure data-sharing platforms. In contrast, individuals increasingly are making their personal information available in social networking and globally, especially in the case of rare diseases, to access improved diagnosis, inform themselves of management and therapy options and to access clinical trials. From the health services perspective, additional safeguards are required to protect confidentiality and patient privacy but without restricting access to improved services, as these new advances facilitate the free flow and sharing of data between different organizations and across international boundaries. In particular, there is a growing concern in the ethics literature and among policy makers that de-identification or coding of personal data and biospecimens is not sufficient for protecting research subjects from privacy invasions and possible breaches of confidentiality due to the possibility of re-identification.[1–3]

As pointed out by Tabor *et al,* some bioethics research and consent discussions tend to focus on identifiability as a harm in itself, rather than on the potential subsequent harms.[4,5] This emphasis may lead to an imbalance between different interests in the ethical assessment where different kinds of potential harms as well as benefits need to be balanced against each other. As argued previously, research subjects have interests at the beginning of the research, for example, on being informed about the purpose, expected benefits and risks of harm, as well as during the process of research, for example, preservation of confidentiality.[6] They also have interests related to the end of the research line, reaping the fruit of research in terms of new, improved and safer medical treatment. Consequently, they have an interest in committing to the widest possible access and most effective use of data and biosamples among researchers as being conducive to those ends. Accordingly, ethical assessments related to data and biobank management in research need to move beyond the narrow focus of identifiability and acknowledge and balance the potential harms and benefits in a broader context. An important primary task then is to explore why it is important to identify individual research subjects, that is, that a given data set refers to one specific individual, entailing the possibility to distinguish between data belonging to different individuals.

AIM

The aim of this paper is to discuss the need to protect the confidentiality of research participants and the need to make them identifiable to facilitate medical treatment and/or for biomedical research.

THE IMPORTANCE OF PROTECTING PRIVACY

The two key elements of privacy are (i) that an individual has access to a secluded private sphere and (ii) that each individual is free to decide who will have access to this sphere, for example, to private information or to a private space.[7] Invasion of privacy can lead to injustice through unfairly discriminatory use of personal information

[1]Department of Public Health and Caring Sciences, Centre for Research Ethics and Bioethics, Uppsala University, Uppsala, Sweden; [2]John Walton Muscular Dystrophy Research Centre, Institute of Genetic Medicine, Newcastle University, Newcastle upon Tyne, UK; [3]Institute of Medical Genetics and Applied Genomics, Rare Disease Center, University of Tübingen, Tübingen, Germany; [4]Division of Pediatric Nephrology, Heidelberg University Center for Pediatric and Adolescent Medicine, Heidelberg, Germany; [5]Department of Neurology, Ulm University Hospital, Ulm, Germany; [6]Office of Rare Diseases Research (ORDR), National Center for Advancing Translational Sciences (NCATS), National Institutes of Health, Bethesda, MD, USA; [7]Office of Population Health Genomics, Public Health and Clinical Services Division, Department of Health Government of Western Australia, Perth, WA, Australia; [8]Centre for Comparative Genomics, Murdoch University, Murdoch, WA, Australia; [9]Centre for Population Health Research, Curtin University of Technology, Bentley, WA, Australia; [10]School of Pathology and Laboratory Medicine, University of Western Australia, Nedlands, WA, Australia; [11]National Center for Rare Diseases, Istituto Superiore di Sanità, Rome, Italy; [12]Institute of rare Diseases research, ISCIII, SpainRDR and CIBERER, Madrid, Spain; [13]Policy Ethics and Life Sciences Research Centre, Newcastle University, Newcastle upon Tyne, UK

*Correspondence: Professor MG Hansson, Department of Public Health and Caring Sciences, Centre for Research Ethics and Bioethics, PO Box 564, Uppsala 751 22, Sweden. Tel: +46 763 41 20 50; Fax: +46 18 471 6675; E-mail: mats.hansson@crb.uu.se
Received 18 August 2015; revised 21 April 2016; accepted 21 April 2016; published online 25 May 2016

though an individual may be harmed merely by having exposed to the public gaze what they would prefer to be private.

Respect for privacy is a means of respecting an individual but it can also be instrumental to establish trust, for example, in medical research contexts. In the context of data sharing, as we discuss in this paper, privacy is to be respected by information and consent procedures and protected by confidentiality procedures. In this paper, we argue that the use of a unique identifier to enable wide data sharing is a proportionate and responsible means of balancing privacy interests with important goods that data sharing is likely to achieve.

THE IMPORTANCE OF IDENTIFYING INDIVIDUALS

In order to be able to accurately diagnose; classify the genetic of rare diseases, and to address the complexity of managing many of these diseases there is a need for good knowledge and a reliable high quality of data. A critical component in building the knowledge and evidence is the need for large volume of clinical data collected and biosamples from affected individuals and their families. These collections enable genotype–phenotype studies to understand how particular genetic changes, that is, the genotype, produce the different disease characteristics we see clinically, that is, the phenotype and how other genes moderate or otherwise the pathology of a disease. This is particularly important in rare diseases where the phenotype results often from a single genetic change. The rarity of the condition is particularly powerful from a clinical perspective because it provides a very high degree of clarity of the genotype–phenotype relationship, and the biological pathways involved; but equally the rarity also represents a barrier to diagnosis and discovery as there are limited data due to the relatively small numbers of affected individuals. Consequently, there is a special and well-documented need for research collaboration so that data and biosamples from multiple independent studies can be shared across borders.[8–10] The two data sets of clinical data and biospecimen data are critical for understanding the pathogenesis and the management of diseases, both common and rare. In relation to empowering clinical diagnosis of diseases and clinical discovery of disease mechanisms, and possible therapies, there are several scenarios where the absence of a re-identification method such as a unique identifier, of data or samples potentially put patients at risk. There is a risk of not reaching scientific objectives in research projects, because data cannot be connected on an individual patient level. Not reaching clinical studies objectives because of insufficient scientific evidence may also put patients at risk by prolonging the period for diagnosis and the options for treatment. Data sets of different experiments and samples of the same patients are held by different research organizations such as samples held at a biobank of university X, clinical or genetic data from a natural history study or registry with hospital Y and biomarker data with research lab Z. Having different organizations involved, including cross border, is a common situation in clinical research for rare diseases. To fully exploit unanticipated, out-of-the normal signals in the biomarker results, the researcher (z) would need to have access to clinical and genetic data (y) of a particular participant from the hospital Y and be in a position to request another sample (x) of the same participant from the biorepository X. In the absence of a re-identification method such as a unique identifier, this research question may remain unanswered. Alternatively, the research needs to be re-done that may not be feasible based on costs and participant/sample availability without access to shared data and resources across borders. In addition, feeding back individual-level results to participants or inclusion in follow-up research may be impossible without a secure way of re-identification.

IDENTIFYING INDIVIDUALS WHILE PROTECTING CONFIDENTIALITY

The basic rationale of a unique identifying system is that a medical researcher puts identifying information about a research participant into a client application that in turn sends encrypted information to a server application, which then returns a generated unique identifier for that individual. A unique identifier for research purposes is a random sequence of characters that is unique to each research participant, regardless of the study, without exposing personally identifiable information. It should also be observed that with such a system 'personal data' in the legal sense is still being processed; and thus, the usual safeguards such as consent and ethics approval must be observed. Identifiers, known as Global Unique Identifiers or GUIDs, for example, NDAR GUID,[11] GRDR GUID (https://grdr.ncats.nih.gov/index.php?option=com_content&view=article&id=113&Itemid=129) and HDI,[12] are meant primarily to facilitate the following patient data in a larger setting. A research network involved in autism research has recently demonstrated how such a system (GUID) may be set up in order to achieve a favourable balance between the competing goals of distinguishing individuals, collecting accurate information for matching and protecting confidentiality.[11] The GRDR GUID facilitates also linking biospecimens and the patient clinical information in GRDR by linking the GUID to the specimen of the same participants (http://ncats.nih.gov/grdr).

A similar system is operated by the European Huntington's Disease Network (EHDN) to generate unique IDs for participants in their studies including the international observational study REGISTRY.[13] Across participating sites, the algorithm for generating that ID is the same. This means that the same ID is generated irrespective of where a given individual takes part in the study. Thus, a participant cannot take part in a study more than once. In addition, the HDID identifier allows the site that enrols the participant to identify that participant at the annual follow-up visit. Further, the same algorithm has been used in other HD studies, for instance TRACK-HD.[12] This HDID identifier means it is possible to merge the data of REGISTRY and TRACK-HD. The same is true for the recently launched global HD observational study Enroll-HD (www.enroll-hd.org). Participants who already took part in REGISTRY keep their unique identifier in Enroll-HD. Once they have consented for Enroll-HD, all data in REGISTRY can be merged with the Enroll-HD database.

Importantly, these systems will generate a unique ID for each participant, if the identifiable information (or the hash code) is correctly transmitted. The system does not retain any personal, identifiable data of the participant, but will recognize whether an ID is requested for the same individual based on the same identifiable data that are required (from the submitter) to generate the ID. We will not go deeper into technical details in this paper, but when setting up systems like this, one has to be sensitive to the possibility that the process of determining which personal identifiers are used to create the unique identification may touch on social, cultural and ethical sensitivities for the research participants. So even though breaches of confidentiality may be avoided there may be other ethical sensitivities to be managed. In the referred autism linkage study, the coordinators experienced significant resistance from subjects in collecting some kinds of personal identifying information, for example, a mother's maiden name and government issued identifier (usually social security number). Upstream consultation with patient organizations can be very useful in facilitating open and informative dialogue, help to identify potential sensitivities and to enable a meaningful informed consent process.

THE RISK OF RE-IDENTIFICATION

There have been several efforts to demonstrate that re-identification of individuals following the sharing of anonymised data is possible by using publicly available databases.[14,15] Gymrek relied on publicly accessible Internet resources in order to show that a combination of a surname with other types of data, such as age and state, can be used to triangulate data and thus reveal the identity of the target.[15] Arguments like these have been used in order to substantiate calls for regulatory changes and policy recommendations.[3,16,17] However, as was shown in a recent systematic review of re-identification attacks (see below) on health data, there are reasons to proceed with caution and to wait for more nuanced accounts before changing policies and regulatory frameworks. It should be remembered, as argued above, that too strict legal regulatory requirements to maintain anonymity may be detrimental to research where maximum benefit from data may only be achieved by distinguishing individuals within the data set.

Searching through 1522 reports of demonstrated re-identifications, El Emam *et al*[18] concluded that the overall success rate for all re-identification attacks was approximately 26 and 34% for health data. However, the confidence interval around these estimates was large, partly because many of the attacks were on small databases. In addition, not all of these examples were using current standards for de-identification, such as the USA Safe Harbor standard or the statistical standard specified in the HIPAA Privacy Rule. We should not be so alarmed whether databases are easily hacked due to negligence regarding the de-identification or coding measures used; as this is to be expected. Of more concern are the serious breaches of existing standards that these databases did not employ the adequate measures to protect participants in the first place; this is the matter for serious concern. El Emam *et al.* found only two studies that succeeded in re-identification when the original data were de-identified in accordance with HIPAA standards. One of these attacks was on health data with the success rate in terms of percentage of records re-identified being 0.013%, which may be considered low. Some of the issues in the policy discussions on re-identification emanate from the deliberate efforts that have been made to identify individuals accessing data associated with queries made through internet services, which are not comparable to the systems described here. When the New York Times reported on the ability to identify a woman it was only possible because her queries included her town name and her personal name (ibid. p.6). It should also be observed that a majority of the demonstration attacks were made by highly qualified experts in the fields of computer science, and they were made for demonstration purposes, not in order to use the information.

El Emam *et al* conclude that it would be prudent for data custodians to use existing standards for de-identification while applying due diligence in prohibiting re-identification attempts as part of data-sharing agreements, providing accountability for data custodian's actions. However, as they assert, this may be appropriate for health data records and databases; regarding genomic information there is need for further analyses where the potential of combining genomic, omic and environmental information for identification of individuals has not yet been sufficiently explored.[19]

RISK OF HARM

A number of potential harms have been associated with the collection and exchange of sensitive personal data. It has been suggested that any re-identification may potentially harm study participants because it will release information on individual disease risks into the public domain.[20] Data subjects might then become vulnerable to the consequences of detrimental genetic information being accessible by insurance companies or employers. Lunshof *et al*[4] suggest that intrusion on privacy can cause harm to social position and opportunities, to personal and family status, for example, identification of an anonymous sperm donor, and to self-image and perception by others. A breach of confidentiality is unlawful, in most jurisdictions, and may also be construed as a personal harm but the issue is rather complex. In Sweden, there is a law requiring consent from a proband in order to contact a genetic relative, for example, to inform about a risk of breast cancer (Law on Genetic Integrity 2006:351). Also, even the referring physician will not get data of genetic testing necessary for further treatment strategies (for instance like cancer or cardiomyopathies) if the patient does not allow it explicitly. Ordinarily, this does not constitute any problem since genetic relatives care for each other or the patient has an interest that his treating physician gets this information and thus allows such contacts. However, there may be exceptions to these moral bonds within a family and if the proband does not allow it, the doctor is bound by law not to tell. In this case breaking the confidentiality agreement may, arguably, lead to decreased risk of harm to others. In the United Kingdom, disclosures are also permissible if the person consents to them but a breach of confidentiality may also be lawful without consent where there is a sufficiently strong public interest at stake thus potentially avoiding a replication of the well-known USA case of Tarasoff in which a psychiatrist failed to breach confidentiality and warn a woman who was killed by one of his patients who had specifically threatened her.[21] From a consequentialist perspective, whether a breach of confidentiality is good or bad is then an open question.

Broken promises and mistrust may also lead to harm. For instance, if a researcher informs an individual study participant that no one will get access to the personal data except the researchers, and the individual later learns that it is not possible to give such a guarantee this may lead to decreased trust and this may in turn lead to harmful consequences, for example, that study participants start opting out of scientific studies. For these reasons, the proposal by Lunshof *et al* that one should not promise more than one can keep and that the rapid development in genetics and bioinformatics calls for an *open consent*, for example, being open to study participants that there is no guarantee of non-identification, was appropriate and timely.[4] They conclude, rightly in our view, that veracity and transparency should be the leading principles in modern research on genotype–phenotype interaction. Thus, research participants must be informed and understand that although all measures will be taken to secure the data and protect the privacy of the individual, there is no 100% guarantee, and there is a small risk that the identity of one individual may be revealed by outsider hackers.

It is common knowledge (personal communication) that patients with rare diseases do acknowledge this risk and are keen to make personal decisions based on their understanding of risks and benefits. Many patients with rare and ultra-rare disease, where a firm clinical or genetic diagnosis is elusive, predictions on inheritance, prognosis and life expectancy cannot be made and treatments are unavailable, consider identification as a minor risk and display their personal information freely on the internet and social media. If children are affected by the rare disease, these decisions will be made by their parents or guardians. In contrast, patients who are pre-symptomatic carriers of a genetic fault that predisposes them to develop a rare condition, but do not show overt signs of the condition and do not have any impact on their life quality, may feel that the risk of identification out-weighs the benefits for research. An important issue

in order to maintain trust and be honest to patients, in our view, is how to consent participants. Consent is a two part process. Part one is providing the participants with all the information about the study and making sure they understate that. For example, participants need clear information about the type, the purpose of the research. The participants must understand that participating in a research does not mean getting treatment. They also need to know why the research is important in finding treatment for them and to other people suffering from the same, related or different diseases. Other examples of information that they need to know.

- How the data will be stored and who the data may be shared with, and what the safeguards for abuse are. This includes information about how likely it is that the Identifier can be hacked.
- What the advantages are of taking part. Be aware that there may not be direct benefits for the participant, but they can help others with the same disease. This is important for the management of expectations.
- What the alternatives are to taking part?
- What costs arise for the participant?
- If data/biosamples can be exploited commercially, and that
- Participation is voluntary.

The second part of the consent is the signature itself, where they sign to agree to participate in the study.

THE NEED TO BALANCE BENEFITS AND HARMS

Hypothetically, it is possible that on first sight, study participants may prefer strong protections of their right to privacy and to autonomy in the sense of making decisions in matters that directly concern them, for example, excluding secondary uses of data and samples without a re-consent, placing strict limits on sharing data and samples or prohibiting unique identifiers. However, on closer inspection, if they are informed about the benefits of sharing and being able to use multiple data and sample collections with the possibility to distinguish individuals, and the costs of re-consenting (or the cost of re-running the research and recollecting the samples and data), they may reconsider.[22] Responsible ethical deliberation need to take into consideration a more comprehensive and broader evaluation balancing the options, both the benefits and costs of acting as well as the benefits and costs of not acting, or of acting in another way. Such broad and comprehensive approaches are not always reflected in the ethics and policy literature on re-identification. As argued by von Wright, to get access of something X that one desires increases one's welfare on the condition that one is informed about the causal relations and consequences both of the totality where X is part as well of the totality where not-X instead of X is part.[23] Informed preferences are more valuable than non-informed, something we will get back to when we later refer to the empirical literature on privacy.

Wrongings in the sense of not being able to keep promises may be prevented by veracity and transparency as leading principles in information and consent procedures. The risk of some harms may be minimized by legislative measures. An essential part of the public trust in medical research using genetic and other kinds of sensitive medical information may depend on patients and research subjects knowing that third parties are prohibited by law from requesting, or inquiring about, genetic or medical information from an individual, with the exception of specified medical situations. This is the case in Sweden. According to Swedish legislation, there has been a shift of attention from putting cumbersome restrictions on research to prevent unauthorized use to making such use in itself unlawful. The law on genetic integrity that came into effect 1 July 2006 laid down that nobody may stipulate as a condition for entering into an agreement, that another party should undergo a genetic examination or submit genetic information about themselves. There should also be a general prohibition to the effect that without support in law, genetic information may not be sought after or used by anyone other than the person that the information is about. This applies even if the person concerned has given his or her consent to such an investigation or use, but not if they themselves have requested it.

The proposed prohibition is not to be applicable to genetic information that is sought for medical purposes, for scientific or genealogical research or to obtain evidence in legal proceedings. For criminal investigations and for insurance purposes, there is regulation in place. Illegitimate requests of or uses of information may still be a problem, but this risk is minimized since such actions will, according to this law, constitute criminal offences. A scale of penalties that includes fines or a term of imprisonment up to 6 months will enforce the proposed prohibitions.[24] Legislation like this and similar legally binding agreements in sharing and access documents may minimize the risk of harm related to unauthorized use. The remaining risks of harm should be balanced against the expected benefits.

HOW CONCERNED ARE PATIENTS AND THE GENERAL PUBLIC?

The literature with concerned scientists, ethicists, lawyers and policy makers is growing almost exponentially in the field of re-identification and privacy. The question is whether these concerns are a true reflection of the potential research participants'? There are a growing number of studies investigating how individuals of different ages look upon and manage risks related to the use of the internet and social media.[25–27] Privacy concerns have also been the focus of many, mostly qualitative studies related to health care and hospital settings.[28–30] Several studies across different medical fields show that patients understand very well that doctors need to share patient information with other doctors involved in their treatment, but this does not include administrative staff.[31,32] The increased use of electronic medical records has inspired a growing number of proposals for technical solutions in order to make information accessible while protecting privacy interests.[33] Issues related to access to medical records for research, for example, in primary health care have been focus for some smaller studies, indicating that patients are not aware about the need for this research but, when told, were concerned about the leakage of information to unauthorized persons.[34] Regarding what kind of information and consent procedure that is appropriate for research using medical records the results are not conclusive. Some want to give explicit consent while others are satisfied with information and the possibility to opt-out.[35–37] Clinical anecdotal evidence indicates that patients with rare diseases are often aware of the fact that there are only limited resources available to carry out research for their particular condition. Therefore, they often altruistically donate their personal time, data and samples to research, and expect that the utility of these gifts is maximized for public good. They are often surprised to hear that data from one research organization cannot be shared with another limiting the conclusions drawn from the research or requiring data and/or samples to be collected again.

Regarding the use of registries and databases with medical and personal information there are only a few studies. A large and well-designed study of values related to this was done in Great Britain in 2005.[38] The background of the study was the perceived increase in regulatory requirements and demands of specific informed consent from ethical review boards and data inspection authorities in relation

to the use of personal data for research, a request made in the name of protection if privacy.[39–41] In all, 2872 individuals in a representative sample of the British population, 97% of those participating in the investigations made by the national statistics bureau in March and April that year, responded to the survey (response rates 62% (March) and 69% (April), respectively). A majority of the British general public did not regard the use of coded data from the national cancer registry for research as an intrusion of privacy. Seventy-two percent did not think that provision of name, address and postal number or receiving a letter with a request for participation in a research project constituted a breach of their privacy. The study draws the conclusion that the increased regulatory requirements are not in accord with what the general public thinks on these matters.

In a study about attitudes to research among participants in a rehabilitation registry in United States, researchers wanted to see whether there were differences in attitudes depending on ethnicity.[42] There was a difference but not that much. In all, 72% of the Afro-Americans and 62% of the white population were positive to participate in registry-based research. As indicated there is need of more studies in order to see whether the protection of privacy suggested in the literature harmonizes with the attitudes of patients and the general public. A deficiency with the current studies that then needs to be acknowledged is that questions are seldom formulated in a way so that one can see how an individual balances different preferences against each other, for example, protection of privacy against the prospect of successful research needed to provide new knowledge and treatment. The need to incorporate different views means that a robust and reproducible method of eliciting preferences for risk of harm is needed: benefit trade-offs must be used to ensure that the resulting values are sufficiently robust to use the information to guide the future development of regulatory policies in this area.

CONCLUSION

Re-identification may potentially bring harmful consequences for the individual, for example, related to insurance and discrimination. On the other hand, as has been demonstrated, there are clear benefits in terms of patient safety with regard to diagnosis and treatment as well as possibilities to generate new and improved treatment through research provided that one can distinguish between individuals by identification. We suggest that open and transparent information and consent processes, not promising too much and the use of safe unique personal identifiers, for example, the GUID or HD identifiers, represent a morally favourable balance, leading to the following recommendations for governance of biobank and data research.

There is no doubt that there are *potential* risks associated with the handling of personal data for medical purposes. The research community must therefore show itself equal to the task of managing those risks and deserving the trust of those whose data it utilizes. The need for a global alliance of responsible data sharing is now well recognized (World Wide Web Consortium), but there must be an ongoing commitment to high professional standards and to the training of the next generation of researchers in the ethics of research.

Adequately informed consent of research participants is one mechanism which contributes to responsible data sharing, but ethical access to, sharing, and use of data are compatible with broad consent, and even no consent, where the justification is ethical, the use lawful and the processing of data is conducted under demonstrably high standards of governance. We have argued for the necessity of using a unique identifier where data from multiple sources are to be accessed, shared and exchanged. In most cases, the use of personal data in research produces benefits which far outweigh the risks though it is important to have an ongoing dialogue with patients and the wider public to ensure that researchers are aware of their concerns and that the public gain an informed understanding of the risks and benefits of research.[43] Researchers should show willing to facilitate upstream engagement with research participants and include governance structures, which reflect the patient/public perspective (could refer to TREAT-NMD or other project examples). As data collection and sharing becomes ever more complex and international there will be a need for harmonized ethical, legal and regulatory approaches and these must recognize the necessary role of a unique identifier to responsible data sharing.

TERMINOLOGY

Identified data

Data labelled or linked to the individual in a way that makes them directly identifiable (name and surname or social security numbers).[44]

***Coded* data (may be single or double coded)**

Personally identifying information is removed and replaced with a code. In the case of *double-coding*, two or more codes are assigned to the same donor's data held in different data sets, with the key connecting the codes back to the donor's direct identifiers held by a third party and not available to the researchers.

Anonymized data

Data that have been identified earlier or coded, but the identification, or the code and the code key have been destroyed, and thus there is no longer any link to the individual.

Anonymous data

There are no links to the individual donor, the data and biospecimens were never associated with identifiers, and the risk of identification of individuals is very low. There may be general descriptions such as 'man, aged 50–55 years, cholesterol level 240 mg per 100 ml'.

Re-identified data

Data that were rendered anonymous but the identity has been retrieved by matching the anonymous information of different databases.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ACKNOWLEDGEMENTS

We acknowledge their involvement in the International Rare Disease Research Consortium (IRDiRC). Funding for this research was received by the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement No. 305444 (RD-Connect) and 305121 (Neuromics). Hansson also received funding by the Innovative Medicines Initiative project BT-Cure (grant agreement number 115142-1), the BioBanking and Molecular Resource Infrastructure of Sweden project (financed by the Swedish Research Council), and the European Union Seventh Framework Programmes Euro-TEAM, BiobankCloud and BBMRI-LPC. Dawkins acknowledges support-in-part from the Australian National Health and Medical Research Council RD-Connect project APP1055319 under the NHMRC–European Union Collaborative Research Grants scheme.

1 Gutman A: Data re-identification: prioritize privacy. *Science* 2013; **339**: 1032.
2 Rodriguez LL, Brooks LD, Greenberg JH, Green ED: The complexities of genomic identifiability. *Science* 2013; **339**: 275–276.
3 Kaye J: The tension between data sharing and the protection of privacy in genomics research. *Annu Rev Genomics Hum Genet* 2012; **13**: 415–431.

4 Lunshof JE, Chadwick R, Vorhaus DB, Church GM: From genetic privacy to open consent. *Nat Rev Genet* 2008; **9**: 406–411.
5 Tabor HK, Berkman BE, Hull SC, Bamshad MJ: Genomics really gets personal: How exome and whole genome sequencing challenge the ethical framework of human genetics research. *Am J Med Genet A* 2011; **155**: 2916–2924.
6 Hansson MG, Simonsson B, Feltelius N, Stjernschantz Forsberg J, Hasford J: Medical registries represent vital patient interests and should not be dismantled by stricter regulation. *Cancer Epidemiol* 2012; **36**: 575–578.
7 Hansson MG The Private Sphere. An emotional territory and its agent, Springer, Philosophical Studies in Contemporary Culture, Monograph, 2008, p 182.
8 Hansson MG, Gattorno M, Stjernschantz Forsberg J, Feltelius N, Martini A, Ruperto N: Ethics bureaucracy – A significant hurdle for collaborative follow-up of drug effectiveness in rare childhood diseases. *Arch Dis Child* 2012; **97**: 561–563.
9 Mascalzoni D, Knopper BM, Ayme S *et al*: Rare diseases and now rare data? *Nat Rev Genet* 2013; **14**: 372.
10 Taruscio D, Gainotti S, Mollo E *et al*: The current situation and needs of rare disease registries in Europe. *Public Health Genomics* 2013; **16**: 288–298.
11 Johnson SB, Whitney G, McAuliffe M *et al*: Using global unique identifiers to link autism collections. *J Am Med Inform Assoc* 2010; **17**: 689–695.
12 Tabrizi SJ, Langbehn DR, Leavitt BR *et al*: TRACK-HD investigators. Biological and clinical manifestations of Huntington's disease in the longitudinal TRACK-HD study: cross-sectional analysis of baseline data. *Lancet Neurol* 2009; **8**: 791–801.
13 Orth M: European Huntington's Disease Network Observing Huntington's disease: the European Huntington's Disease Network's REGISTRY. *J Neurol Neurosurg Psychiatry* 2011; **82**: 1409–1412.
14 McGuire A, Gibbs RA: No longer de-identified. *Science* 2006; **312**: 370–371.
15 Gymrek M, McGuire AL, Golan D, Halperin E, Erlich Y: Identifying personal genomes by surname inference. *Science* 2013; **339**: 321–324.
16 Rothstein MA: Is deidentification sufficient to protect health privacy in research? *Am J Bioeth* 2010; **10**: 3–11.
17 Ogbogu U, Burningham S, Ollenberger *et al*: Policy recommendations for addressing privacy challenges associated with cell-based research and interventions. *BMC Med Ethics* 2014; **15**: 7.
18 El Emam K, Jonker E, Arbuckle L, Malin B: A systematic review of re-identification attacks on health data. *PLoS One* 2011; **6**: e28071.
19 Malin B, Loukides G, Benitez K, Clayton E: Identifiability in biobanks: models, measures and mitigation strategies. *Hum Genet* 2011; **130**: 383–392.
20 Wjst M: Caught you: threats to confidentiality due to the public release of large-scale genetic data sets. *BMC Med Ethics* 2010; **11**: 21.
21 Lowenthal D: Case studies in confidentiality. *J Psychiatr Pract* 2002; **8**: 151–159.
22 Hansson MG: For the safety and benefit of current and future patients. *Pathobiology* 2007; **74**: 198–205.
23 von Wright, Georg, HenrikThe good of man. In: Carson, Thomas L, Moser, Paul K (eds.): *Morality and the Good Life*. Oxford University Press: New York, NY, USA, 1997, pp.147–163.
24 Lag om genetisk integritet m.m. (Law on Genetic Integrity), 2006: 351.
25 Boyd D, Marwick AE. Social Privacy in Networked Publics: Teens' Attitudes, Practices, and Strategies. A Decade in Internet Time: Symposium on the Dynamics of the Internet and Society, September 2011. Available at: http://ssrn.com/abstract=1925128 (accessed 22 September 2011).
26 Davis K, James C: Tween's conceptions of privacy online: implications for educators. *Learn Media Technol* 2013; **38**: 4–25.
27 Merolli M, Gray K, Martin-Sanchez F: Health outcomes and related effects of using social media in chronic disease management: a literature review and analysis of affordances. *J Biomed Informatics* 2013; **46**: 957–969.
28 Karro J, Dent AW, Farish S: Patient perceptions of privacy infringements in an emergency department. *Emerg Med Australas* 2005; **17**: 117–123.
29 Walsh KI: Nurses' and patients' perceptions of dignity. *Int J Nurs Pract* 2002; **8**: 143–151.
30 Malcolm HA: Does privacy matter? Former patients discuss their percpetions of privacy in shared hospital rooms. *Nurs Ethics* 2005; **12**: 156–166.
31 Schers H, van den Hoogen H, Grol R, van den Bosch W: Continuity of information in general practice. Patient views on confidentiality. *Scand J Prim Health Care* 2003; **21**: 21–26.
32 Whiddett R, Hunter I, Engelbrecht J, Handy J: Patients' attitudes towards sharing their health information. *Int J Med Informatics* 2006; **75**: 530–541.
33 Fernández-Alemán JL, Carrión Senor I, Lozoya PAO, Toval A: Security and privacy in electronic health records: a systematic literature review. *J Biomed Informatics* 2013; **46**: 541–562.
34 Clerkin P, Buckley BS, Murphy AW, MacFarlane AE: Patients' views about the use of their personal information from general practice medical records in health research: a qualitative study in Ireland. *Fam Pract* 2013; **30**: 105–112.
35 Stevenson F, Lloyd N, Harrington L, Wallace P: Use of electronic patient records for research: views of patient and staff in general practice. *Fam Pract* 2013; **30**: 227–232.
36 Buckley BS, Murphy AW, MacFarlane AE: Public attitudes to the use in research of personal health information from general practitioners' records: a survey of the Irish general public. *J Med Ethics* 2011; **37**: 50–55.
37 Willison DJ, Steeves V, Charles C *et al*: Consent for use of personal information for health research: do people with potentially stigmatizing health conditions and the general public differ in their opinions? *BMC Med Ethics* 2009; **10**: 10.
38 Barret G, Cassel JA, Peacock JL, Coleman MP: National survey of British public's views on use of identifiable medical data by the national cancer registry. *Br Med J* 2006; **332**: 1068–1072.
39 Iversen A, Liddell K, Fear N, Hotopf M, Wessely S Consent, confidentiality and the Data Protection Act. *BMJ* 2006; **332**: 165–169.
40 Ward HJ, Cousens SN, Smith-Bathgate B *et al*: Obstacles to conducting epidemiological research in the UK general population. *BMJ* 2004; **329**: 277–279.
41 Hansson MG: Do we need a wider view of autonomy in epidemiological research? *Br Med J* 2010; **340**: c2335, 1172–1174.
42 Phipps E, Harris D, Brown N *et al*: Investigation of ethnic differences in willingness to enroll in a rehabilitation research registry. *Am J Phys Med Rehabil* 2004; **83**: 875–883.
43 Academy of Medical Sciences: *Personal Data for Public Good: Using Health Information in Medical Research*. AMS: London, 2006.
44 Mascalzoni D, Dove E, Rubinstein Y *et al*: International charter of principles for sharing bio-specimens and data. *Eur J Hum Genet* 2014; **23**: 721–728.