Special Section: Open Forum

Communicating Identifiability Risks to Biobank Donors

T.J. KASPERBAUER, MICKEY GJERRIS, GUNHILD WALDEMAR, and PETER SANDØE

Abstract: Recent highly publicized privacy breaches in healthcare and genomics research have led many to question whether current standards of data protection are adequate. Improvements in de-identification techniques, combined with pervasive data sharing, have increased the likelihood that external parties can track individuals across multiple databases. This article focuses on the communication of identifiability risks in the process of obtaining consent for donation and research. Most ethical discussions of identifiability risks have focused on the severity of the risk and how it might be mitigated, and what precisely is at stake in pervasive data sharing. However, there has been little discussion of whether and how to communicate the risk to potential donors. We review the ethical arguments behind favoring different types of risk communication in the consent process, and outline how identifiability concerns can be incorporated into either a detailed or a simplified method of communicating risks during the consent process.

Introduction

An issue of increasing concern in biobanking is the unintended or unauthorized release of personal information about biobank donors. The standard tool for protecting donor data is anonymization, which is intended to strip data of personally identifying information, while still providing important information for researchers and clinicians. However, a combination of factors has caused some to question whether standard anonymization techniques are adequate for protecting donors against retrieval of their personal information by third parties.

One major risk arises out of pervasive data collection both within and outside of biobanking and healthcare contexts. It has proven possible to link supposedly anonymous data to specific individuals by comparing information across multiple databases.[1] Improvements in genetic analysis have facilitated this sort of identification, increasing risks not only to individuals but to their close relatives as well.[2] These risk factors are amplified by widespread sharing of databases internationally, making it difficult to know who is responsible for regulating the data and ensuring anonymization.[3] Finally, many biobanks store material and data indefinitely, for currently unforeseen purposes, which entail corresponding unforeseen identifiability risks.[4]

Most ethical discussions of these identification risks have focused on the severity of the risk and how it might be mitigated, and what precisely is at stake in pervasive data sharing. However, so far it has not been much discussed whether and how to communicate the risk to potential donors.

In the context of biobanking for stem cell research, Ubaka Ogbogu et al. report that it is standard practice to communicate privacy risks to potential donors, although donors are not always asked to explicitly consent to these risks.[5] There have, however, been high-profile breaches of privacy, particularly in genomics research, where it turned out that participants were not informed of potential privacy or identifiability risks.[6] This has led some to question the way that risks are commonly communicated in biobank research. Deborah Mascalzoni et al., for

Cambridge Quarterly of Healthcare Ethics (2018), **27**, 123–136.

doi:10.1017/S0963180117000457

example, argue that data sharing and de-anonymization risks are often overlooked in the design of consent procedures.[7] Paul Ohm furthermore argues that standard consent procedures place the burden entirely on donors to understand an extremely complex issue, such that it is unlikely that donors sufficiently comprehend the relevant concerns.[8] Others, however, have argued that no particular change is needed to the consent process to communicate identifiability risks. Lisa Parker, for example, argues that most biobank research projects do not pose significant identifiability risks, and, therefore, do not require donor consent. What is more important, she argues, is proper oversight and ethical review of specific biobank projects.[9]

In light of these debates, our goal in this article is to outline and discuss ethical arguments relating to whether and how to communicate identifiability risks, as part of responsible biobank management. Thereby we hope to fill some gaps in the literature concerning why and how to secure consent from donors.

Biobanks and Identifiability

The Practice of Biobanking

Our focus here is on biobanks used for storing human biological samples as well as associated health and personal information. There are two main types of human biobanks relevant to our investigation.[10] The first are biobanks that routinely collect samples for unspecified or general purposes. For example, hospitals routinely collect blood, skin, and other bodily fluids for unspecified future clinical use and for research. There are also national biobanks in many countries that function as a general resource for research aiming at improving population health. National biobanks typically aim to collect blood samples in order to provide an accurate representation of the population, or particular subgroups of the population.

Second, many biobanks collect samples relevant to specific diseases or disorders, or for specific research purposes. For example, the Danish Dementia Biobank in Denmark collects samples from patients who are being treated for various neurodegenerative diseases. Similar biobanks exist in many countries for cancer, heart disease, and myriad other diseases. Some biobanks also store blood or tissue from specific organs. Biological samples can also be collected for short-term research projects, which are then disposed of at the completion of the project.

Privacy and identifiability issues have arisen out of both types of biobanking, and in both clinical and research settings. However, the most contentious debates arguably occur in biobanking for research purposes, where personal information is more widely shared than in the clinic. This is partly because of recently created massively collaborative biobank networks, which are designed to provide open access to researchers. For example, the EuroBioBank Network and the Biobanking and Biomolecular Resources Research Infrastructure (BBMRI) connect researchers in Europe.[11] These networks improve the quality of research, especially into rare diseases, but also complicate data management. There are also initiatives to create stronger links between large biobank databases and electronic health records, in order to improve patient care.[12] These too exacerbate privacy issues with biobanks, as we discuss subsequently.

Our discussion is meant to apply to massively collaborative biobanking as well as to relatively mundane collections of biological samples. The issues we discuss

encompass biobanks used for either general or specific purposes, and for both research and the clinic. We will return to biobank policies on privacy and related issues after further discussing the relationship among anonymity, privacy, and identifiability.

Anonymity, Privacy, and Identifiability

The concepts of anonymity, privacy, and identifiability are closely related. We focus on identifiability in this article because we see it as raising a number of concerns with regard to data sharing. To clarify our position, we will explain how we understand these terms and what we see as fundamentally important about identifiability in biobanks.

Our understanding of privacy and anonymity draws from recent work by Jeffrey Skopek. He argues that privacy and anonymity should be viewed as complementary to each other, such that "Privacy involves hiding the *information*, whereas anonymity involves hiding what makes it *personal*."[13] In other words, privacy refers to limiting access to people's information, whereas anonymity, by contrast, refers to concealing whose information it is.[14]

For example, suppose that someone who has donated tissue to a biobank has also been diagnosed with the early stages of Alzheimer's disease. Others who access this biobank's data may be able to see clinical notes, including the Alzheimer's diagnosis. One common view, according to Skopek, is that we use anonymization in this context in order to *protect* sensitive information. If the Alzheimer's diagnosis is considered private, we could protect privacy by removing identifying information; for example, by identifying the patient with a number instead of the person's name (also known as pseudonymization, as discussed subsequently).[15]

Skopek, however, thinks this is misleading. Even if the information is anonymized, it is accessible by others, and, therefore, is no longer absolutely private. Since anonymization does not limit access to information, it should not be understood as a method for protecting privacy.

This way of distinguishing privacy and anonymity can be debated, but we think this distinction helps to illustrate the fundamental importance of identifiability. Identifiability, as we understand it, refers to tracking specific individuals or groups of individuals; linking them, for example, to sensitive information about their health. Therefore, protections against identifiability aim at securing the anonymity of the patient.

Common pieces of information included in biobank records (and healthcare records generally) are birth dates, race, zip codes (or regional identifiers), sex, and disease information. Marital status and information about offspring and family members are also sometimes included. Some form of identification is used as well, but this is usually anonymized. Anonymization aims to remove direct identifiers such as names and any other piece of information that is directly tied to personal identify (such as national identification numbers). Pseudonymization replaces direct identifiers with indirect identifiers, such as random sequences of numbers.[16]

However, the absence of names, as many have pointed out, does not mean the absence of identifiability. Any of these pieces of information could be used to facilitate identification of specific individuals, if they are sufficiently unique across multiple databases. Consider, for example, someone who is identified only by sex, birth date, and having been diagnosed with Alzheimer's disease. Depending on the

database, the Alzheimer's diagnosis could be a unique identifier, especially if it is a rare familial form of Alzheimer's. There may only be one person in the database who is, for example, a 47-year-old woman with Alzheimer's disease. Someone with access to multiple databases where this particular patient's data is held can use this information to gather additional information. Suppose that in another database she is only identified as having Alzheimer's disease and living in Texas. If the Alzheimer's diagnosis is rare in both databases, it increases the reliability of inferring that the 47-year-old woman from the first database also lives in Texas. This is a highly stylized example, but it illustrates the basic phenomenon as well as the importance of identifiability.

The fundamental issue is that the risk of identification increases as our personal information is widely shared across multiple databases. Even when our information remains relatively private (e.g., when shared only in healthcare databases), it often carries unique identifiers. As we will explain, personal information can sometimes be connected to individuals by comparing multiple databases carrying unique identifiers. This is not a problem with loss of privacy as such, however. Sharing private information with one's physician, for example, is not by itself an issue. Rather, the issue is the combination of pervasive information sharing and recent developments in identification techniques (we return to these factors in the context of DNA in the section entitled "Risks to biobanks and health information"). Taken together, these developments raise significant risks to identifiability in the context of biobanking.

Politicians, civil servants, and researchers involved in managing and regulating biobanks are well aware of these developments. However, very little advice has been forthcoming on how these new risks should impact the process of informed consent. There is a strong presumption in favor of communicating privacy risks in all major international guidelines on biobanking and data sharing. The Organisation for Economic Cooperation and Development's (OECD) 1980 guidelines on privacy, which have been widely influential, require that patients and donors be notified of the privacy policy protecting their information, and that consent be obtained indicating that donors agree to those terms.[17] These guidelines were updated in 2013, and although the updated version identifies identifiability risks as an issue of concern, no specific advice is offered.[18] The OECD's 2009 guidelines specifically on biobanks suggest that consent forms include "The general procedures and safeguards used to protect privacy and confidentiality," as well as whether material or data might be shared with third parties, including law enforcement, commercial entities, insurers, and employers.[19] These are presented only as suggestions, however, not fundamental requirements.

Other recently proposed international guidelines advance much stronger requirements for communicating identifiability risks. The 2016 recommendations from the World Health Organization (WHO) and the Council for International Organization of Medical Sciences (CIOMS) state, "During the process of obtaining informed consent, those responsible for the biobank must inform the potential donors about the safeguards that will be taken to protect confidentiality as well as their limitations."[20] They further specify that "Donors must be informed of the limits to the ability of researchers to ensure strict confidentiality and of the potential adverse consequences of breaches of confidentiality." The limitations they mention include accidental leaks or stolen data, targeted attacks using re-identification techniques, and the possibility that information sharing might be required for legal purposes.

Similar considerations are included in the World Medical Association's (WMA) 2016 guidelines.[21]

Therefore, although there is a presumption in favor of communicating identifiability risks, there is disagreement about the number and type of identifiability risks that must be communicated, as well as the amount of details that should be communicated. In order to better assess how identifiability risks should be communicated to biobank donors, we must first determine the exact nature of the risk.

Current Threats to De-identification

Risks of Re-identification

Recital 26 of the 1995 European Union Data Directive states that "to determine whether a person is identifiable, account should be taken of all the means likely reasonably to be used either by the controller or by any other person to identify the said person."[22] This phrase "likely reasonably" has provided a framework for discussing the severity of recent threats to de-identification.[23] The Data Directive suggests that single cases or hypothetical cases do not compromise protection against identification; they are not sufficiently likely. However, many have expressed concern about identification risks based on such cases, especially where they indicate the presence of general deficiencies in data protection.

Paul Ohm has forcefully argued that current forms of data collection and storage are inherently risky, and that anonymization is inadequate for protecting against identification.[24] Ohm argues that the main problem is re-identification, or the ability to identify individuals by comparing information across multiple separate databases. The Article 29 Data Protection Working Party, which analyzes privacy and identifiability issues arising out of the European Union Data Directive, identifies three main forms of re-identification: (1) *singling out*, or identifying specific people, even if not by name, (2) *linkability*, or identifying groups of individuals, even if specific individuals cannot be identified, and (3) *inference*, or deducing traits based on information in a database.[25]

Each of these methods exploits the fact that even when anonymized, databases often carry unique information about individuals. Consider again the Alzheimer's diagnosis mentioned. If an individual's Alzheimer's diagnosis is rare across multiple databases, it becomes much easier to glean other information about that individual from those databases (singling out), just as it is if there is a small group of people with that diagnosis (linkability). Even if the diagnosis is not mentioned in the database, other information may indicate that people in the database have received an Alzheimer's diagnosis (inference). For example, a database containing information about Internet activity could include search terms concerning Alzheimer's, disease, or perhaps even Alzheimer's-disease-related medicine that has been purchased online.

According to Ohm, there may for each of us be a "database of ruin" that possesses the right combination of information that uniquely identifies us, thereby enabling repeated releases of potentially embarrassing private information. "Accretive reidentification," he says, "makes all of our secrets fundamentally easier to discover and reveal."[26] To support this claim, Ohm discusses three prominent cases of re-identification: internet usage by AOL users, the Massachusetts Governor's health

data, and movie rankings by Netflix users.[27] In each of these cases, widely available anonymized databases were compared in order to find unique identifiers.

For illustration, consider the governor of Massachusetts, who was made vulnerable by the release of information about all Massachusetts state employees' hospital visits. This information was anonymized and made freely available to researchers. The database contained information about sex, zip codes, and birth dates, the combination of which is known to uniquely identify large portions of the American population. Although data controllers are now more aware of these unique identifiers, similar identifiers may exist in other widely available databases.

Ohm and others have taken these cases to illustrate the vulnerability of large anonymized databases. Some, however, have questioned whether these cases indicate the presence of significant risks. Jane Yakowitz, for, example, argues, "the risks imposed on data subjects by datasets that do go through adequate anonymization procedures are trivially small."[28] Yakowtiz characterizes the current state of data collection and storage as a "state of highly unlikely risk."[29]

There are three important points that Yakowitz makes in support of the insignificance of these risks. First, cases such as those mentioned are uncommon, and do not reflect general deficiencies in data protection. Although they were highly publicized, and did indeed expose surprising gaps in data protection, normal protocols are arguably adequate in most cases.[30] For example, she cites studies also cited by Ohm indicating that the Health Insurance Portability and Accountability Act (HIPAA) in the United States is largely effective at preventing re-identification using health data (and would have prevented the Massachusetts governor case). Second, Yakowitz argues that the techniques required to successfully circumvent anonymization are highly technical, and unavailable to most people. Access to many databases is very expensive, and even widely available databases require sophisticated analytical knowledge in order to extract useful information. Third, she suggests that determined attackers have easier routes for gaining information than investing in re-identification techniques. Scanning blog posts, for example, is easier and may provide much more useful information than trying to infer information about individuals by inspecting anonymized databases.

These considerations suggest that the risk of re-identification is low. In the terminology of the risk assessment framework used for chemicals, the *hazard* is high because health information carries unique identifiers, but the *exposure* is low because exploiting unique identifiers is difficult.[31] Yakowitz concludes that large anonymized databases containing personal information are no more risky than our garbage. It is true that others can use it to access private information, potentially providing unique identifiers; however, future re-identification is unlikely, and whatever information is gathered likely will not be that damaging. To combat the risk, health professionals should continue to employ anonymization and other methods to make identification difficult, and should also prevent against accidental releases of information, which Yakowitz thinks is indeed risky, but beyond that, there is no particular concern with large databases containing personal information.

In response to Yakowitz, others have argued that prominent cases of re-identification provide a "proof of principle" that has altered the data collection landscape. The abovementioned cases sparked a debate among cryptographers, for example, about the adequacy of de-identification techniques, given the apparent deficiencies with anonymization.[32] They disagree about the extent of the in-principle risk of re-identification; for example, whether any method exists that can successfully

defend against very determined and skilled attackers. But there is general consensus that many databases fail to employ the best available methods (a claim Yakowitz also agrees with). Current standards of data collection and sharing are low risk *only if* data controllers employ the right protective methods.

The appropriate methods vary, however, according to the type of database and the purpose of the data collection. To obtain a more precise estimate of the relevant risks, we must, therefore, address the risks specific to biobanks and personal health information.

Risks to Biobanks and Health Information

A handful of studies have recently been published on health information breaches. Perhaps the most comprehensive comes from the Nuffield Council's Working Party on Biological and Health Data.[33] They reviewed European Union and United Kingdom legal databases between 1995 and 2014 (from LexisNexis and the United Kingdom Information Commissioner's Office) to find evidence of leaked health information, as well as breaches discussed in newspapers and on Twitter. The legal databases revealed 36 cases in the United Kingdom, and another 14 in the European Union more broadly. They also found 87 cases mentioned in newspapers and another 70 mentioned on Twitter. The evidence for these was less systematic ("soft evidence," as they called it), however, so we will focus on the legal cases.

The breaches documented in the legal databases were analyzed for their causes as well as for the resulting harm (according to many legal definitions of harm). They determined that the most common cause of information breaches (10 of the 51 cases across the United Kingdom and European Union) was administrative mistakes (e.g., failure to follow correct procedures), followed by explicit sharing of information against the individual's wishes (9 cases), and human error (7 cases). Four of the 51 cases were attributed to "insufficient safeguards," in which the data protection procedures themselves were deficient.

They also evaluated the documented harms with all 51 information breaches. They did so by determining whether there was evidence of "emotional or physical, individual distress," a common legal definition of harm. Eighteen of the 51 cases met this criterion, while another 27 were determined to carry the potential for harm (the remaining 6 were considered harmless).

None of the cases analyzed involved biobanks, nor were there instances of targeted attacks using the advanced re-identification techniques discussed. These cases raise similar issues, however, about the vulnerability of de-identified personal health information; 51 cases over the course of 19 years might seem insignificant; however, these were only the most thoroughly documented (enough for legal proceedings). Moreover, there were indeed cases in which the protocols either were not followed or did not provide adequate protection. Therefore, it would seem that health data carry some risk even in the absence of targeted attacks using sophisticated technology.

In the United States, any health information breach involving more than 500 records must be reported to Congress. From 2009 to 2014, 1,187 such breaches were recorded, affecting more than 41,000,000 people.[34] These numbers suggest that health information in the United States is highly vulnerable.

The only systematic analysis of such breaches comes from El Emam et al.[35] They reviewed the available evidence (in 2010) of successful re-identification in data

sets that had undergone de-identification procedures. They found 14 cases, 6 of which involved health information (all in either the United States or Canada). Across those six health-related databases, an estimated 34 percent of the records could be re-identified. It was further determined that only one database out of the six had fully implemented adequate measures against de-identification (according to HIPAA guidelines). Within that database, however, only 2 out of 15,000 records could be re-identified.

This evidence suggests that health information is vulnerable. A 34 percent success rate from targeted attacks on de-identified data sets does indeed seem to be a significant risk. This type of attack supports the "proof of principle" idea mentioned. However, the evidence also indicates that de-identification measures, when appropriately implemented, are effective. If HIPAA standards (or something like them) had been followed, the targeted attacks would likely have been much less successful.

Another type of risk that has been widely discussed in relation to biobanks comes from genomic databases. The in-principle risks are arguably higher with biobanks designed specifically for genomics analyses, because genetic information makes certain inferences across databases easier. Yaniv Erlich and Arvind Narayanan's review of identification breaches in genomics databases identifies certain classic techniques, such as using birth dates and zip codes to identify participants in the Human Genome Project.[36] But they also review cases in which, for example, the disposition for Alzheimer's disease could be inferred in close family members. Although no re-identification occurred to those family members, it was shown to be possible in principle. Similarly, Suyash Shringapure and Carlos Bustamente found that it is possible to identify specific individuals in large genomics databases that are searchable through "beacon" websites, which only allow yes or no questions about single nucleotides found in the database.[37] A program designed to ask repeated yes or no questions was able to identify specific individuals within several thousand pointed questions. Some beacon websites index nonpublic information, such as medical diagnoses. This facilitates inferences between individuals and their family members.

This is pertinent to biobanks because the biological material most important to biobanks contains DNA. Mark Taylor has argued that, from a privacy perspective, biological material in a biobank should be treated as genetic data, because their "interpretive potential" is the same.[38] Biobanks typically store material for long periods of time, and future accessibility is often uncertain; therefore, biobank material carries the potential to be reanalyzed in much the same way as genomic information. As a result, relevant risks also apply to family members of those who participate in biobanks. This is especially important with familial diseases, particularly if donors or their family members do not want anyone else to be informed about the disease, including family members who may not yet be aware of the disease. Therefore, even though there have been relatively few identification breaches with biobanks, there is significant risk because of the number of people potentially impacted.

The risk to biobanks is amplified by widespread sharing of information from biobanks, especially across jurisdictions.[39] Edward Dove identifies biobanking as one of the main areas in which sharing data internationally has increased risks to privacy.[40] The relevant regulations, he argues, are less precise and effective than with local control. Similarly, Harald Schmidt and Shawneequa Callier argue that

identifiability often changes as biological data changes hands, and that legal protections often stipulate a definition of identifiability that applies only to specific (and temporary) circumstances.[41] All of these authors note that although there have not been many data breaches to date, there is also currently no oversight. Without better oversight, attempted breaches are hard to detect or prevent.

Returning to the European Union Data Directive, it appears that a significant portion of identification risks are sufficiently "likely reasonable" to demand regulatory action. We turn now to how these risks should be communicated to biobank donors.

A Framework for Communicating Identifiability Risks

As discussed in the section entitled "*Anonymity, privacy, and identifiability*," there is a strong presumption in favor of communicating privacy risks in all major international guidelines on biobanking and data sharing. However, the OECD's guidelines on both privacy and biobanking offer limited guidance on communicating identifiability concerns, as do the guidelines from the WHO/CIOMS and the WMA. There is a presumption in favor of communicating identifiability risks, but little agreement about the number and types of identifiability risks that should be communicated, and few details about the number of details that should be communicated. Here, we review the ethical reasons behind favoring different types of risk communication in the consent process, and outline how identifiability concerns can be incorporated into either a detailed or a simplified method of communicating risks during the consent process.

Limited and Simple Communication of Identifiability Risks

The main reason usually cited for the importance of obtaining consent is that it is essential for preserving donor/patient autonomy.[42] The WHO/CIOMS guideline mentioned previously states "Informed consent protects the individual's freedom of choice and respects the individual's autonomy." Autonomy is also expressed as the basis of informed consent in the Helsinki Declaration, the Belmont Report, and (to a lesser extent) the Nuremberg Code. Control is central to this conception of autonomy. By asking donors for their informed consent, donors are allowed to decide whether the risks and burdens of donation are acceptable. Having the choice to accept these risks grants donors some control over the use of their biological material.

The conditions that must be met to preserve autonomy are far from clear, however, particularly with respect to communicating risks. Ruth Faden and Tom Beauchamp's classic *A History and Theory of Informed Consent* argues that "substantial understanding" of foreseeable consequences and possible outcomes is required in order to preserve autonomy,[43] but as they and many others have pointed out, substantial understanding might be accomplished best by simplifying the nature of the risk when communicating with donors and patients.

It might be objected that simplifying risk communication clearly undermines autonomy. Omitting details in the consent process is usually justified only if there are clear benefits that outweigh the autonomy of donors and patients.[44] However, simplicity does not necessarily entail deception or inaccuracy. On the contrary, some have argued that simplified risk communication *enhances* donor autonomy.

Identifiability risks are so complex that thorough and detailed risk communication may fail to provide adequate comprehension. Donors are likely to simplify the information themselves, but in ways incompatible with the nature of the risk. The OECD's privacy guidelines nicely summarize this problem: "Individuals tend to rely on "rules of thumb" when making decisions, a tendency that may lead them to ignore certain options or simply not make a choice. They also present inconsistencies when weighing probabilities, and may appear to place more value on the present than on the future. In turn, such behaviours affect how information is absorbed. More information for individuals about an organisation's privacy practices and personal data usage may not always be better."[45]

Extensive and detailed communication might thus hamper donors' understanding of the relevant risks. Simplified language, by contrast, can increase autonomy because it is more readily understood, thereby helping donors make informed choices about the use of their material.

What might simplified communication of identifiability risks look like? When considering the identification risks discussed, the following points could be applied.

- The details of indirect identification are arguably too difficult to comprehend, and would need to be omitted.
- The potential for targeted attacks and accidental leaks are difficult to communicate simply, and may distract donors from the more general point that privacy and anonymity cannot be guaranteed.
- Donors could be informed that privacy and anonymity cannot be guaranteed, even if they do not receive an explanation for why.
- Donors could also be notified that it is difficult to predict how personal information might be shared in the future.
- The inherent identifiability of DNA could perhaps also be formulated in plain language.

This level of risk communication would also likely omit any mention of statistics or studies indicating the probability of the risks.

Simplified risk communication seems preferable, especially when the relevant identifiability risks are particularly low. Lisa Parker argues that most identifiability concerns in biobanking are sufficiently unlikely that they are best dealt with by ethics committee review, rather than individual consent.[46] A reasonable alternative, however, is to frame these risks in simple terms, and to limit the number of potential risks identified. In cases of low risk, it may also be helpful to compare the risk to other types of data collection and sharing in healthcare contexts. In many countries (e.g., the United States), many types of routine health data collection (e.g., sharing patient data among hospitals) receive minimal consent, if at all, and pose similar identifiability risks to biobanking. This sort of comparison would presumably aid in comprehension.

Extensive and Detailed Communication of Identifiability Risks

It is widely accepted that simplistic communication can sometimes aid in comprehension. However, many have argued that extensive and detailed communication of risks is nonetheless preferable. How might one argue, contrary to what was discussed, that identifiability risks must be communicated in detail?

Solon Barocas and Helen Nissenbaum discuss a "transparency paradox" with informed consent: Clear and simple language is required for donors to comprehend the relevant risks, but is not sufficiently detailed or precise to produce truly *informed* consent.[47] They argue, "For individuals to make considered decisions about privacy in this environment, they need to be informed about the types of information being collected, with whom it is shared, under what constraints, and for what purposes."[48] Plain or general language about identification risks just is not sufficient. This presents extra difficulties for those obtaining consent, but without these details, the consent process is, according to the two authors, meaningless.

Details are particularly important, they emphasize, because of uncertainties about future data sharing and future possible identifiability risks. It is possible that data sharing policies will change in the future, and the donors themselves might be unavailable to re-obtain consent. De-identification might also become easier in the future. Donors must, therefore, be informed that current protections could become inadequate. Consent forms in genomics research usually emphasize that privacy cannot be guaranteed, given the inherent identifiability of DNA, and that the samples are stored for long periods of time.[49] Perhaps this should also be required for consent in biobanking.

Ohm makes a similar point about the meaninglessness of simplified risk communication in the consent process.[50] Merely notifying people of potential risks (e.g., that unintended identification could occur, with no additional details) fails to provide an adequate basis for comprehending the content of one's consent. Identifiability risks formulated in general language are too easily dismissed, leading people to consent without understanding the hazards that they might face in the future.

Risks that meet a certain threshold of probability and significance may, therefore, need to be communicated in detail. This would provide sufficiently informed consent for protecting autonomy. What might detailed communication of identifiability risks look like? Considering again the preceding list, we envision that details would need to be added to each of the following points.

- Indirect identification, including details about singling out, linkability, and inference
- Targeted attacks
- Accidental leaks of personal information
- The presence of data in multiple databases
- Pervasive data sharing, both domestically and internationally
- Long-term storage of biomaterial and associated data
- Identifiability through DNA analysis

All of these risks would need to be identified and communicated to donors, in addition to a number of other details about the source and extent of the risk, including statistics indicating their probability, when possible. For example, the studies mentioned previously about the extent of the harm caused by identification would help give donors an idea about the significance of the risk. Risks to family members would also need to be outlined, as would mitigation steps and contingency plans, should their personal information be leaked. Also relevant would be details about data storage and handling, as well as the extraction of genomic information from biomaterial.

Conclusion

We have discussed possible ways of communicating identifiability risks, and outlined important ethical concerns in choosing to communicate identifiability risks in either simplified or more detailed formats. This is just a first step toward integrating these concerns into current consent practices. Much more work is needed to determine what is required for different types of biobanks, depending on the services they provide.

As discussed, biobanks are typically distinguished by use for either general or specific purposes. Another relevant distinguishing feature is that some biobanks store pseudonymized, rather than fully anonymized samples. Anonymization is counterproductive if material is kept in order to provide personalized treatment, if follow-up data will be needed, or if donors must be re-contacted in the future (e.g., in order to obtain consent for secondary uses of their material). Communicating identifiability risks in these cases is important because pseudonymized material is easier to track. Detailed communication may therefore be more pertinent, because individuals are potentially more susceptible to indirect identification. For large national biobanks, especially those that regularly share information with researchers, individuals may also be more susceptible to accidental leaks and targeted attacks. If these events are sufficiently likely, detailed communication during the consent process would seem appropriate.

Although anonymization might help protect against identification, anonymization is sometimes understood as permitting secondary uses, even without donor consent (which is partly what has motivated changes to the Common Rule in the United States). This places a greater burden on proper communication of relevant risks during the initial consent process. It can also be the case that more information is shared about an individual for use in research than would be shared for routine biobanking (e.g., to learn more about a rare disease). This too might call for more detailed risk communication, considering the potential for future releases of information and the inability to re-obtain consent.

Finding the balance between different degrees of communication to obtain truly informed consent in different contexts and at the same time striking a balance between enabling research and protecting the autonomy of donors, raises a number of ethical challenges. With increased generation, storing, and sharing of health data, these challenges will only grow.

Notes

1. Barbaro M, Zeller T, Jr. A face is exposed for AOL searcher no. 4417749. *New York Times*, August 9, 2006, at A1; Sweeney L. Uniqueness of simple demographics in the U.S. population. Laboratory for International Data Privacy Working Paper, 2000; Narayanan A, Shmatikov V. Robust deanonymization of large sparse datasets. *IEEE Symposium on Security and Privacy* 2008;8:111–25.
2. Lowrance WW, Collins FS. Identifiability in genomic research. *Science* 2007;317:600–602; Wjst M. Caught you: Threats to confidentiality due to the public release of large-scale genetic data sets. *BMC Medical Ethics* 2010;11:21–4; Erlich Y, Narayanan A. Routes for breaching and protecting genetic privacy. *Nature Reviews Genetics* 2014;15:409–21; Shringapure SS, Bustamente CD. Privacy risks from genomic data-sharing beacons. *American Journal of Human Genetics* 2015;97: 631–46.
3. Dove ES. Biobanks, data sharing, and the drive for a global privacy governance framework. *Journal of Law, Medicine, & Ethics* 2015;43:675–89.
4. Solomon Cargill S. Biobanking and the abandonment of informed consent: An ethical imperative. *Public Health Ethics* 2016;9:255–63.

5. Ogbogu U, Burninham S, Ollenberger A, Calder K, Du L, El Emam K, et al. Policy recommendations for addressing privacy challenges associated with cell-based research and interventions. *BMC Medical Ethics* 2014;15:1–7.
6. Jacobs KB, Yeager M, Wacholder S, Craig D, Kraft P, Hunter DJ, et al. A new statistic and its power to infer membership in a genome-wide association study using genotype frequencies. *Nature Genetics* 2009;42:1253–7.
7. Mascalzoni D, Hicks A, Pramstaller P, Wjst M. Informed consent in the genomics era. *PLoS Medicine* 2008;5:1302–5.
8. Ohm P. Changing the rules: General principles for data use and analysis. In Lane J, Stodden V, Bender S, Nissenbaum H, eds. *Privacy, Big Data, and the Public Good Frameworks for Engagement*. Cambridge: Cambridge University Press; 2014:96–111.
9. Parker L. Using human tissue: When do we need consent? *Journal of Medical Ethics* 2011;37: 759–61.
10. Marko–Varga G, Baker MS, Boja ES, Rodriguez H, Fehniger TE. Biorepository regulatory frameworks: Building parallel resources that both promote scientific investigation and protect human subjects, *Journal of Proteome Research* 2014;13:5319–24; Hewitt RE. Biobanking: The foundation of personalized medicine. *Current Opinion in Oncology* 2011;23:112–9.
11. Mora M, Angelici C, Bignami F, Bodin A-M, Crimi M, Di Donato J-H, et al. The EuroBioBank Network: Ten years of hands-on experience of collaborative, translational biobanking for rare diseases. *European Journal of Human Genetics*. 2015;23:1116–23.
12. Scott CT, Caulfield T, Borgelt E, Illes J. Personal medicine—The new banking crisis. *Nature Biotechnology* 2012;30:141–7.
13. Skopek JM. Reasonable expectations of anonymity. *Virginia Law Review* 2015;101:694.
14. Also see Skopek JM. Anonymity, the production of goods, and institutional design. *Fordham Law Review* 2014;82:1751–1809.
15. For a discussion of different interpretations of anonymization, see Schmidt H, Callier S. How anonymous is 'anonymous'? Some suggestions towards a coherent universal coding system for genetic samples. *Journal of Medical Ethics* 2012;38:304–9.
16. If a third party holds the key-code connecting an identifying number to the patient, the patient's information is considered pseudonymized. If no key-code exists, the information is considered completely anonymized. We group pseudonymization and anonymization together here because identifiability issues apply equally to both.
17. OECD. Guidelines on the Protection of Privacy and Transborder Flows of Personal Data, 2013; available at http://www.oecd.org/sti/ieconomy/oecd_privacy_framework.pdf (last accessed 25 Aug 2017).
18. Similarly, the European Union General Data Protection Regulation, which will go into effect May 25, 2018, requires explicit consent whenever personal data is collected, including notification of the purposes for which the information will be used, but does not specify the communication of identifiability risks. Available at http://eur-lex.europa.eu/legal-content/EN/TXT/?uri=uriserv:OJ.L_.2016.119.01.0001.01.ENG (last accessed 25 Aug 2017).
19. OECD. Guidelines on Human Biobanks and Genetic Research Databases, 2009; available at https://www.oecd.org/sti/biotech/44054609.pdf. Accessed 08/25/2017.
20. Council for International Organizations of Medical Sciences (CIOMS) and the World Health Organization (WHO). International Ethical Guidelines for Health-Related Research Involving Human Subjects, 2016, at 44; available at https://cioms.ch/wp-content/uploads/2017/01/WEB-CIOMS-EthicalGuidelines.pdf (last accessed 25 Aug 2017).
21. WMA. Declaration of Taipei on Ethical Considerations Regarding Health Databases and Biobanks, 2016; available at https://www.wma.net/policies-post/wma-declaration-of-taipei-on-ethical-considerations-regarding-health-databases-and-biobanks/ (last accessed 25 Aug 2017).
22. Directive 95/46/EC of the European Parliament and of the Council of October 24, 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data. Available at http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:31995L0046:en:HTML (last accessed 25 Aug 2017).
23. For an attempt to model identifiability risks, see Malin B, Loukides G, Benitez K, Clayton EW. Identifiability in biobanks: Models, measures, and mitigation strategies. *Human Genetics* 2011; 130:383–92.
24. Ohm P. Broken promises of privacy: Responding to the surprising failure of anonymization. *UCLA Law Review* 2010;57:1701–77.

25. Article 29 Data Protection Working Party. Opinion 05/2014 on anonymisation techniques, April 10, 2014; available at http://www.cnpd.public.lu/fr/publications/groupe-art29/wp216_en.pdf (last accessed 25 Aug 2017).
26. See note 24, Ohm 2010, at 1705.
27. See note 1, Barbaro, Zeller 2006, at A1; Sweeney 2000; Narayanan, Shmatikov 2008.
28. Yakowitz J. Tragedy of the data commons. *Harvard Journal of Law & Technology* 2011;25, at 37.
29. See note 28, Yakowitz 2011, at 40.
30. There are a number of strategies that data controllers employ to protect against such attacks. For a review, see El Amam K. *Guide to the De-Identification of Personal Health Information*. Boca Raton, FL: CRC Press; 2013.
31. van Leeuwen CJ, Vermeire TG. *Risk Assessment of Chemicals. An Introduction*. Dordrecht: Springer; 2007.
32. See, for example, the debate initiated in Cavoukian A, Castro D. Big data and innovation, setting the record straight: De- identification does work, 2014; available at http://www2.itif.org/2014-big-data-deidentification.pdf (last accessed 25 Aug 2017); Narayanan A, Felten EW. No silver bullet: De-identification still doesn't work, 2014; available at http://randomwalker.info/publications/no-silver-bullet-de-identification.pdf (last accessed 25 Aug 2017); El Emam K, Arbuckle L. De-identification: A critical debate, 2014; available at https://fpf.org/2014/07/24/de-identification-a-critical-debate/ (last accessed 25 Aug 2017).
33. Laurie G, Jones KH, Stevens L, Dobbs C. A review of evidence relating to harm resulting from uses of health and biomedical data. Nuffield Council on Bioethics Working Party on Biological and Health Data 2014; available at http://nuffieldbioethics.org/wp-content/uploads/FINAL-Report-on-Harms-Arising-from-Use-of-Health-and-Biomedical-Data-30-JUNE-2014.pdf (last accessed 25 Aug 2017).
34. Office for Civil Rights. *Annual Report to Congress on Breaches of Unsecured Protected Health Information for Calendar Years 2013 and 2014*. Washington, DC: United States Department of Health and Human Services; 2015.
35. El Emam K., Jonker F, Arbuckle L, Malin B. A systematic review of re-identification attacks on health data. *PLoS One*, 2011;6:e28071.
36. See note 2, Erlich, Narayanan 2014.
37. See note 2, Shringapure, Bustamente 2015.
38. Taylor M. *Genetic Data and the Law: A Critical Perspective on Privacy Protection*. Cambridge: Cambridge University Press; 2012.
39. See note 2, Wjst 2010.
40. See note 3, Dove 2015.
41. See note 15, Schmidt H, Callier 2012.
42. Beauchamp T. Autonomy and consent. In Miller F, Wertheimer A, eds. *The Ethics of Consent: Theory and Practice*. Oxford: Oxford University Press; 2009:55–74.
43. Faden R, Beauchamp T. *A History and Theory of Informed Consent*. Oxford: Oxford University Press; 1986, at 278.
44. See note 5, Ogbogu et al. 2014.
45. See note 17, OECD 2013.
46. See note 9, Parker 2011.
47. Barocas S, Nissenbaum H. Big data's end run around anonymity and consent. In: Lane J, Stodden V, Bender S, Nissenbaum H, eds. *Privacy, Big Data, and the Public Good: Frameworks for Engagement*. Cambridge: Cambridge University Press; 2014:44–75.
48. See note 47, Barocas, Nissenbaum 2014, at 58.
49. McGuire AL, Besko LM. Informed consent in genomics and genetic research. *Annual Review of Genomics & Human Genetics* 2010;11:361–81.
50. See note 8, Ohm 2014.