Published in final edited form as:
Hum Genet. 2011 September ; 130(3): 383–392. doi:10.1007/s00439-011-1042-5.

Identifiability in biobanks: models, measures, and mitigation strategies

Bradley Malin,
Department of Biomedical Informatics, School of Medicine, Vanderbilt University, 2525 West End Avenue, Suite 600, Nashville, TN 37203, USA. Department of Electrical Engineering and Computer Science, School of Engineering, Vanderbilt University, Nashville, USA

Grigorios Loukides,
Department of Biomedical Informatics, School of Medicine, Vanderbilt University, 2525 West End Avenue, Suite 600, Nashville, TN 37203, USA

Kathleen Benitez, and
Department of Biomedical Informatics, School of Medicine, Vanderbilt University, 2525 West End Avenue, Suite 600, Nashville, TN 37203, USA

Ellen Wright Clayton
Department of Pediatrics, School of Medicine, Vanderbilt, USA. Center for Biomedical Ethics and Society, School of Medicine, Vanderbilt University, 2525 West End Avenue, Suite 400, Nashville, TN 37203, USA. School of Law, Vanderbilt University, Nashville, USA

Bradley Malin: b.malin@vanderbilt.edu; Grigorios Loukides: gloukides@acm.org; Kathleen Benitez: kathleen.benitez@vanderbilt.edu; Ellen Wright Clayton: ellen.clayton@vanderbilt.edu

Abstract

The collection and sharing of person-specific biospecimens has raised significant questions regarding privacy. In particular, the question of *identifiability*, or the degree to which materials stored in biobanks can be linked to the name of the individuals from which they were derived, is under scrutiny. The goal of this paper is to review the extent to which biospecimens and affiliated data can be designated as identifiable. To achieve this goal, we summarize recent research in identifiability assessment for DNA sequence data, as well as associated demographic and clinical data, shared via biobanks. We demonstrate the variability of the degree of risk, the factors that contribute to this variation, and potential ways to mitigate and manage such risk. Finally, we discuss the policy implications of these findings, particularly as they pertain to biobank security and access policies. We situate our review in the context of real data sharing scenarios and biorepositories.

Introduction

The medical community is in the midst of a personalization revolution that promises to make healthcare more efficient, effective, and safe (Collins 2010; Glaser et al. 2008; Hamburg and Collins 2010). It is anticipated that one of the major contributors to this revolution will be molecular medicine, where systems biology and genomics are leading the charge (Burke and Psaty 2007; Green et al. 2011; Ng et al. 2009; Roses 2004). While the notion of a biorepository is not a new concept [i.e., the medical community has collected and stored biospecimens for centuries (Eiseman et al. 2003)], the decreasing cost of high-

Correspondence to: Bradley Malin, b.malin@vanderbilt.edu.

throughput technologies, combined with recent advances in information technology in the clinical setting, has set the stage for large-scale biomedical association mining and translational discoveries (Bellazi and Zupan 2008; Ritchie et al. 2010). In support of these activities, organizations across the country and around the globe are stockpiling biospecimens to facilitate medical research (e.g., Ollier et al. 2005; Louie et al. 2007). In particular, a growing number of healthcare institutions are integrating biorepositories with data derived directly from the clinical setting (e.g., Kullo et al. 2010; Lemke et al. 2010; Roden et al. 2008). For instance, the NIH-sponsored electronic medical records and genomics (eMERGE) network consists of a consortium of medical centers utilizing biobanks to perform genome–phenome association studies with clinical phenotypes derived from medical information systems applied in primary care environments (McCarty et al. 2011).

Until recently, the collection, analysis, and application of clinical and genomic information were localized to specific investigators or institutions. Increasingly, however, scientists are urged and at times required to share data to strengthen the statistical power of complex association experiments and to allow the research community to replicate and verify clinically relevant findings [e.g. (Guttmacher and Collins 2005; National Institutes of Health 2003, 2007)]. To assist scientists in achieving these goals, agencies around the globe continue to invest considerable effort to construct information technology infrastructure, such as the Database of Genotype and Phenotype (dbGaP) at the US National Library of Medicine (Mailman et al. 2007), which will facilitate the consolidation, standardization, and dissemination of patient-specific records from disparate investigators. Other countries, such as the UK, which created the UK Biobank, have chosen to centralize primary collection of data (Ollier et al. 2005).

At the same time, the increased collection and sharing of sensitive biomedical information have raised significant societal issues, including concerns over patient privacy, which could easily derail these efforts (McGuire and Gibbs 2006). One of the major privacy issues has been *identifiability*, i.e., the extent to which materials and data stored in biobanks can be linked to the name of the individuals from which they were derived. The goal of this review paper is to determine the extent to which biospecimens, and derived data can be designated as identifiable.

Identifiability and data sharing policies

In general, one of the primary strategies that organizations have traditionally used to deal with privacy threats is by defining and adhering to data sharing policies.[1] For instance, when sharing of data from NIH sponsored investigations is required, the NIH policies (National Institutes of Health 2003, 2007) specify that the data should be disseminated in a manner that is devoid of identifiers. In practice, the NIH recommends the removal of an enumerated list of potential identifiers, similar to the “Safe Harbor” de-identification standard of the Privacy Role associated with the Health Insurance Portability and Accountability Act of 1996 (U.S. Department of Health and Human Services 2002). This list includes explicit identifiers, such as names and Social Security Numbers, as well as potential quasi-identifiers, such as dates and geocodes more specific than the first three digits of a zip code. Yet, when after such features are removed from clinical information associated with genomic data, the residual sequence of nucleotides can be well distinguished. But, it should be recognized that the ability to distinguish records, whether genomic or clinical, from each other is not the same as the ability to identify from whom they came, a point that we expand upon below. The statistics are beyond the scope of this paper, but we note that, by some

[1] Other strategies leverage security measures to limit access, but are beyond the scope of this discussion. We refer the reader to Langella et al. (2008) and Lemrow et al. (2007) for further discussions on such security practices.

estimates, only about 100 single nucleotide polymorphisms (SNP) are required to distinguish an individual's DNA record (Lin et al. 2004, 2006). Moreover, it may be possible to ascertain information about the genetic or clinical status (if it is included in the record) of family members based on the heritability relations (Cassa et al. 2008). It is also possible to ascertain ancestral origin (Phillips et al. 2007) and some investigators are exploring ways to infer broad physiognomic characteristics from genomic sequence (e.g., Kayser and Schneider 2009; Ossorio 2006). As a result, until recently, it was the policy of dbGaP to publicly post online only the aggregate case–control information for each SNP in a study (i.e., the likelihood a person from the case group harbored a particular SNP variant, and similarly for the control group). Concerns were subsequently raised, however, over reports that, even when an individual's DNA is disseminated in an aggregated form, an individual with knowledge of a particular person's DNA could determine if he or she was in the case group, control group, or neither group (Clayton 2010; Homer et al. 2008; Wang et al. 2009). In response, the NIH and Wellcome Trust removed genomic summaries of case and control cohorts from the public section of databanks, including dbGaP (Zerhouni and Nabel 2008).

Certainly, such attacks on patients' privacy are plausible, but the ability of perpetrators to utilize genomic data to compromise privacy is, for the time being, limited. The main reason is that perpetrators (i.e., the people seeking to identify an individual in a dataset) must possess an identified reference sample of DNA, which typically is hard to come by. In addition, it begs the question of how likely such an attack can be performed. Later in the paper, we posit scenario in which the DNA records in a biobank may be exploited for identification purposes; however, at this point we wish to impress upon the reader that a greater risk resides in the possibility of matching clinical records with public information (El Emam 2008; Lowrance and Collins 2007). This point cannot be overstated. For instance, in the 1990s, it was famously illustrated that one could purchase the Cambridge, MA voter registration list for $20 and link it to a public version of the state's hospital discharge database through the combination of *date of birth*, *gender*, and *residential zip code*, thus revealing the identities associated with many clinical diagnoses, including the governor's diagnosis (Sweeney 1997). This event provided impetus for the Safe Harbor standard mentioned earlier.

We urge that defining the effectiveness of de-identification strategies at reducing re-identification risks is necessary to develop ethically sound research policy. Using well-characterized tools for de-identification will promote more informed choice and thereby encourage the inclusion of data from a broader array of people than "information altruists" who disclaim specifically any guarantee of anonymity (Kohane and Altman 2005; Lunshof et al. 2008). Knowledge of the actual re-identification risk associated with a given dataset also would help resolve whether data are underprotected and in need of additional safeguards, or overprotected such that data sharing policies could be more permissive.

A risk-based framework to identifiability

In addition to the case mentioned above, an increasing number of investigations demonstrate how genomic and health information, devoid of explicit identifiers, could be re-identified to the corresponding patient (e.g., El Emam et al. 2006; Loukides et al. 2010a; Malin and Sweeney 2004; Sweeney 1997). However, it is important to recognize that there is a significant difference between the description of a path by which such information could be re-identified and the likelihood that such a path would be leveraged by an adversary in the real world (Malin et al. 2010). In this regard, regulations such as HIPAA Privacy Rule, are not specified in a manner that precludes the dissemination of data that could be re-identified. Rather, the Privacy Rule explicitly states that the extent to which health information can be designated as de-identified must account for the context of the anticipated recipients who use reasonable means to attempt to re-identify the information.

As such, we should consider the broader environment in terms of how a *reasonable* recipient would attempt to pursue re-identification. Table 1 summarizes the principles that could be utilized to determine if health data are sufficiently de-identified (Malin et al. 2010). These principles build on those defined by the Federal Committee on Statistical Methodology [which is referenced in the original publication of the Privacy Rule, see (Subcommittee on Disclosure Limitation Methodology, Federal Committee on Statistical Methodology 2005)]. In general, it helps to separate the health information attributes, or types of data, into classes of relatively “high” and “low” risks. Although risk actually is more of a continuum, this rough partition illustrates how context impacts risk.

Based on the criteria described in Table 1, we can now perform a *risk assessment*. The greater the replicability, availability, and distinguishability of the health information, the greater the risk for re-identification. As an example of a low risk environment, consider that laboratory values may be very distinguishing, but they are rarely independently replicable and are rarely disclosed in multiple and widely accessibly resources. In contrast, as an example of a high risk environment, consider that demographics can be highly distinguishing, are highly replicable, and are available in public resources.

Modeling and Measuring Re-identification Risks

For illustration purposes, we apply this framework to a known, highly likely threat. Particularly, we focus on recent work which provides a demonstration of how decision makers can model and measure privacy risks in an easy to digest manner with respect to existing policies in the context of known threats (Benitez and Malin 2010). We aim to use this forum to illustrate the power and insight such an approach can provide to decision makers. As a starting point, it is crucial to recognize that, when disparate organizations adhere to the same data sharing policy, the privacy risks will vary mainly because the organizations function in varying regulatory contexts and manage data on different populations. The question one must now ask is: once information is shared, to what extent can someone with little knowledge (apart from the shared data and other public resources) exploit it for re-identification purposes? The general format of a re-identification attack is depicted in Fig. 1.

In this case, a *de-identified dataset* released from an information holder, such as a medical facility, is found to have commonalities with another dataset drawn from the same population. The latter dataset contains names or other identifying information and it is known as the *identified dataset*. If a record in the de-identified dataset matches only one of the records in the identified dataset, there is a potential for unique re-identification. Such matches are certain re-identifications only if the identified dataset contains information on everyone in the de-identified dataset, otherwise there may be one or more individuals who have the same characteristics over the attributes used to perform the match but are not represented in the identified sample. The more complete the identified dataset, and the more fields in common between the two datasets, the less the likelihood of false re-identifications.

A recent study investigated risks associated with re-identification attacks that require nothing more than a computer, some data, and a basic knowledge of spreadsheets or databases in personal computing software, such as Microsoft Office (Benitez and Malin 2010). We derived a baseline estimation of re-identification achieved through demographics using the US Census and statistical estimation techniques (Golle 2006). Data from the US Census were selected because it provides robust estimates of the US population, and thus the number of unique persons based on the demographic attributes constitutes a ceiling on the number of true re-identifications possible through such attributes, namely *county*, *date of birth*, *gender*, and *race*. In Texas in the year 2000, for instance, approximately 31,000 people were estimated to be unique based on these four attributes. In contrast, approximately 300 people in Delaware were estimated as unique based on the same characteristics. These numbers correspond to 0.14 and 0.03% of the total population in the year 2000 for Texas and Delaware, respectively. These estimates are a measure of re-identification potential, but such demographics may not be available in identified datasets because datasets are often subject to some sort of policy-based transformation before they are shared.

To paint a more complete picture of the effects that policies exert on re-identification risk, the study investigated two data sharing policies currently in use, both set forth in the Privacy Rule of the Health Insurance Portability and Accountability Act (HIPAA): Safe Harbor and Limited Dataset (U.S. Department of Health and Human Services 2002). Data released under Safe Harbor provisions are approved for distribution to a wide audience. To conform to Safe Harbor, 18 potentially identifying features must be removed from the data prior to its dissemination. These features include dates and geographic areas with a population smaller than 20,000 people, for example. The Safe Harbor policy permits the disclosure of demographic features such as race, gender, residential state, and year of birth.[2] In contrast, information shared according to the Limited Dataset can contain additional, more detailed data, including dates and zip codes. However, recipients of data that are shared according to the latter policy must sign a use agreement that prohibits re-identification. That said, it can be estimated how much more risky it is to disseminate records using the Limited Dataset in comparison to Safe Harbor policy. Using Tennessee as an example, Benitez and Malin (2010) estimated that there were approximately 60 unique people in the state (0.001% of the population) who are vulnerable to re-identification, based on the demographics available in a release that consists of race, gender, residential state, and year of birth. In contrast, they estimated 1.8 million people in Tennessee (32% of the population) are unique based on demographic information permissible in Limited Dataset releases. This constitutes a 30,000-fold increase in re-identification risk, which suggests that the drafters of HIPAA trusted researchers 30,000 times more than the general public. This trust differential multiplier varies from state to state, ranging from less than 1,000 to more than 100,000, roughly increasing with the size of a state’s population. Clearly, the same types of data released in two different states are subject to two very different levels of privacy risk.

From Distinguishable to Identified

The final piece of the puzzle is the identified dataset. Each state, and each record holder within that state, has different guidelines or regulations on handling and release of public records. Voter registration lists, mentioned in the earlier re-identification attack by Sweeney (1997), are not the only kind of public record, nor are public records the only sources of identified demographic information. However, voter registration lists are considered an ideal identified dataset mainly because they:

- cover a large portion of the adult population,
- generally contain current information, typically contain a wide variety of information, and
- can be obtained at low cost.

[2] It should be noted that Safe Harbor actually permits the first 3-digit zip code of a region to be disclosed when the population is greater than 20,000. We use the simplification of state of residence for illustrative purposes and because it has been observed that many organizations choose to withhold such information in their application of the policy.

To characterize the threat of voter registration databases in today's climate, the previous mentioned identifiability study surveyed the elections department of each US state (Benitez and Malin 2010). It was found that current policies differ widely on many dimensions, including who is eligible to receive copies of voter registration data, what information is included, and how much the lists cost. For instance, 30 states included information on voters' year of birth, while only six included information about race.

Given this knowledge, re-identification risk estimates were recalculated with respect to the availability of states' voter registration databases. When permissive Limited Dataset stipulations were docked against the information available in the public records, it was observed that the number of fields available for re-identification tends to decrease, as does the risk.

In some states, such as Tennessee, for instance, the risk does not change significantly. Overall, however, 45 states revealed less information in their public voter registration lists than was available through Limited Dataset provisions. In many states, such as Oklahoma, the risk was slightly lower than the original Limited Dataset estimate. In other states, such as Ohio, the risk approached that of the Safe Harbor policy. In certain extreme cases, such as Wisconsin, very little demographic information was available in voter rosters that the re-identification threat virtually disappeared. Intruders attempting to re-identify data from such states would be hampered, not by the health data protection policy, but by the paucity of identified information available from their state.

Thus, while current data sharing policies seek to create a level playing field of privacy risk, the landscape is more complex and varied than even the example just discussed. There are myriad types of public datasets, as well as re-identification attacks (Malin 2005a) and as the amount of data grows, the likelihood of successful attacks may increase.

Estimating the number of people with a unique combination of features is applicable to non-demographic attacks as well. Finding naming sources for certain types of data will be more difficult than others, but assuming that such sources exist, it is desirable to have some estimate of the distribution for the values found in the intersecting fields.

Towards risk mitigation

We should not be content with measuring risk, but should proactively mitigate it. This can be achieved using an array of strategies that have been developed by federal statistical agencies to protect survey data and have collectively been referred to as *statistical disclosure control* approaches. These strategies were designed to generate data that preserve certain aggregate statistics, without revealing the data of any particular individual and include *noise addition* (e.g., random value changes in a record), *data swapping* (e.g., exchanging values across records) and *synthetic data generation* (e.g., data based on properties of the original records, without corresponding to any real individuals). A proper survey of these methods is beyond the scope of this paper, but we direct the reader to several excellent surveys on the topic (Adam and Wortman 1989; Willenborg and De Waal 1996).

While offering solid privacy guarantees, the majority of such methods have been of limited application for data deposited to biorepositories. A primary reason is that they can ascribe to individuals values they did not originally have. Thus, risk mitigation strategies specialized to health and genomic data tend to focus on strategies that are able to preserve *data truthfulness*. Two popular methods that address this requirement are *generalization*, which replaces values with more general but semantically consistent values, and *suppression*, which removes values from the released data (Sweeney 2002a; Samarati 2001; Bayardo and Agrawal 2005; El Emam and Dankar 2008).

However, these methods should not be applied in an ad hoc manner because, if not used properly, they can overdistort or inappropriately protect records. Rather, they are often applied in the context of formal anonymization models, such as what we term *k*-based models. These models are based on the premise that each record must be indistinguishable from at least $k - 1$ other records with respect to quasi-identifiers. Variants of these models include *k*-map (Sweeney 2002b)*, k*-anonymity (Sweeney 2002a; Samarati 2001), privacy-constrained anonymity (Loukides et al. 2010b), *k*-unlinkability (Malin 2007, 2008), and *k*-ambiguity (Vinterbo et al. 2001). Without delving too far into the details of these models, it is important to recognize that they differ in the assumptions made about a data recipient's ability to leverage an identified source for re-identification purposes. For instance, *k*-map assumes that an attacker attempts to link each published record to the entire population from which a patient was derived. More concretely, if a record in a biorepository was submitted by Vanderbilt University Medical Center, then the recipient may assume that the corresponding individual was from the surrounding vicinity, and might be any resident of Tennessee, Kentucky, or Alabama. Alternatively, the *k*-anonymity model assumes that the recipient is more knowledgeable and is aware of the exact set of people for which the records correspond. For example, if a cohort consists of 100 patients, it is assumed that the recipient knows who the 100 patients are, but not which exactly their record is nor their genomic sequences.

Approaches that enforce *k*-based models using generalization and suppression have been applied to various types of data that could be exploited for re-identification purposes. For instance, they have been utilized to protect patient demographics (Chiang et al. 2003; El Emam and Dankar 2008; El Emam et al. 2009; Sweeney 2002a; Vinterbo et al. 2001, Wang et al. 2004), genome sequences (Lin et al. 2002; Malin 2005b; Li et al. 2011), and diagnosis codes such as International Classification of Disease codes (Loukides et al. 2010a, b). At the same time, it is crucial to recognize that these approaches must be adapted for the type of data they are applied to and the intended purposes of use.

For illustration, let us take a moment to expand on the diagnosis code anonymization problem. The attack involves the use of an identified dataset containing individuals' names and diagnosis codes, which can be obtained in several ways: (1) by accessing a health care provider's electronic medical record system from which the de-identified data has been derived, or (2) by combining public records (e.g., voter registration lists) with de-identified hospital discharge records. By linking the identified to the de-identified datasets, based on the combinations of *potentially distinguishing* diagnosis codes, an attacker can associate individuals with their de-identified records to infer their diagnoses, as well as sensitive information, such as genomic sequences contained in samples that are disseminated in the context of a genome wide association study (GWAS). Guarding against such an attack while ensuring that the released dataset permits the discovery, and validation, of clinically useful associations between diagnoses and genomic variants is a challenging computational task. To achieve this task, Loukides et al. (2010b) proposed a method to group potentially distinguishing diagnosis codes together to satisfy a *k*-based model, which requires each published record to be equivalent to at least $k-1$ other records in the published dataset with respect to these diagnosis codes. The ability of this method to produce data that both mitigates re-identification attacks and remains useful for conducting GWAS was empirically shown using a real cohort of patients' records that were to be deposited in dbGaP. Specifically, using a sample of approximately 3,000 patients for whom a GWAS was run on native electrical conduction within the ventricles of the heart, it was shown that diagnosis codes indicative of various cancers with known SNP associations, such as breast cancer, lung cancer, and pancreatic cancer, could be shared without violating the formal privacy model.

Though *k*-based models afford a provable level of protection for each record, they are stricter than is required by regulations and may hinder data analytics (Benitez et al. 2010). One reason is that these models set the risk of a data-set to that of the least protected record, thereby assuming a worst-case scenario. As an alternative, a model based on the average risk has been proposed (Benitez et al. 2010; Dankar and El Emam 2010) and evaluated on a variety of patient cohorts submitted by the eMERGE network to dbGaP (Malin et al. 2011). Through this exercise it was illustrated that solutions could be tailored to the needs of the cohorts. For instance, it was illustrated that detailed age information on a cohort of elderly patients involved in a dementia GWAS could be disclosed provided certain demographics were generalized, such as rare ethnicities with minimal ability to contribute sufficient power to association studies. We further note that a risk-based model that is similar in principle to those of Benitez et al. 2010 and Dankar and El Emam 2010 but not developed for guarding against re-identification, has been proposed by Sankararaman et al. 2009. The model attempts to prevent an attacker from determining if an individual is characterized as a case, control, or neither of the two. It is applied to pooled DNA sequence data (i.e. individual SNP vs. phenotype status) and is useful in determining how many, and which, SNPs could be shared publicly.

Table 2 summarizes four popular approaches based on generalization and suppression that have been evaluated on the types of data that may be disseminated into biorepositories, namely demographics and standardized diagnosis codes. For each approach, we discuss the re-identification attack it addresses, the privacy principle and transformation method it applies, and how it attempts to ensure the data remain useful for biomedical and genetic analysis.

While the existing re-identification mitigation approaches are an important step forward, enabling data providers to measure the re-identification risk of the data they intend to disseminate remains challenging. A critical issue at this juncture is that there are no agreed upon standards of acceptable level of re-identification risks.

Public concerns: how realistic?

Despite the evidence above that re-identification is largely preventable, some members of the public remain worried about the use of research data by people outside the health care and research enterprises to identify individuals (Botkin 2001; Clayton et al. 2010; Haga and O'Daniel 2011; McGuire et al. 2008b). Concerns include insurance and employment discrimination (Lemke et al. 2010), paternity identification (Miler 2009) and, in particular, use by the criminal justice system (Lemke et al. 2010). Clinical DNA databases can be used for forensic purposes, as it was dramatically illustrated by the case of Anna Lindh, the former Swedish Minister for Foreign Affairs who was stabbed to death in a Stockholm department store. The police obtained the newborn blood spot of the alleged murderer to confirm the murderer's identity and elicit a confession (Hansson and Björkman 2006). A subsequent survey of the Swedish public reported that more than 85% thought it was acceptable for police to access these kinds of samples for criminal investigations (Bexelius et al. 2007). Similar requests for identified clinical data sets have been made in other jurisdictions as well (Hindmarsh and Abu-Bakar 2007).

Not all observers or members of the general public have been so supportive of forensic use of clinical and research samples (Mccartney 2004; Kaye 2006). The NIH has long allowed local investigators and their institutions to obtain certificates of confidentiality to protect identified research data from forced disclosure, evincing a clear policy choice that research trumped the justice system (National Institutes of Health 2002). These certificates, however, are not widely used, and questions have been raised about their effectiveness despite occasional anecdotes of their utility (Currie 2005; Wolf and Zandecki 2006). In any event, these certificates are not available for databanks located within the federal government, such as dbGaP.

Adults in the US appear to be concerned about forensic uses. Recent focus groups asking almost 5,000 adults in the United States their views about participating in a hypothetical de-identified national biobank reported that “84% felt that it would be important to have a law protecting research information from law-enforcement officials” (Kaufman et al. 2009). Despite these concerns, the NIH in its most recent iteration of the GWAS data sharing policy, which governs dbGaP, “acknowledges that legitimate requests for access to data made by law enforcement offices to the NIH may be fulfilled” (National Institutes of Health 2007). But rather than launching a wholesale assault on the GWAS data sharing policy, the more relevant question may be to ask how likely it is that law enforcement would try to access data in dbGaP or any other de-identified biobank for purposes of identifying a person. The answer depends not only on the difficulty of re-identifying someone using these de-identified datasets but also the likely availability of other sources of more readily identified data, such as CODIS (Anonymous 2011) and obtaining DNA samples from relatives (Miller 2010). The resulting risk that law enforcement would seek access to dbGaP to try to identify a criminal is almost surely quite low.

A final note

Much has been made of the uniqueness of an individual’s DNA sequence, but it is not yet possible to identify a person without an identified sample of DNA. Science is simply not good enough at present, and it probably will never be, to predict complete phenotype from genomic DNA, save for some Mendelian traits. Nonetheless, there are certain manners by which DNA in biorepositories can be exploited for identification purposes. As such, biobanking managers and policy managers should keep in mind the following points when addressing identifiability issues.

- Recognize the Difference Between Perceived and Realistic Risks: The literature reports on numerous ways by which data stored in biorespositories could be re-identified. Yet, many of these exploits require significant effort and luck to accomplish successfully. For instance, one of the more likely risks to realize is that someone who has a sample of DNA from an identified individual would seek to determine whether that person’s DNA was in a research dataset. However, it is difficult to imagine why someone would want to do that, except to prove that it is possible, or at least why making such a match would cause any harm to the individual. In this vein, we recommend that managers be vigilant regarding the difference between proof of concept attacks published in the literature, as well as reported on in the media, versus realistic attacks. One will never create a perfectly secure system or, in this case, a system devoid of re-identification risks.
- Build Realistic Models of Identifiability: Beyond uniqueness of a DNA sequence, additional data types in biorepositories (e.g., demographics, standardized codes) could be leveraged for re-identification purposes. We encourage managers to be vigilant and model which features could be leveraged for identification purposes, through which resources, and by whom.
- Quantify and Mitigate the Identification Risks: Once practical models of risk are defined, biorepository managers adopt appropriate approaches to measure re-identification risk. It is important to measure risk for their specific repository because risks are context-dependent. The risk associated with a de-identification policy for a particular repository does not necessarily transfer to another repository due to differences in patient populations, availability of identified resources, and

cost associated with perpetrating an attack. In additional, when risks are determined to be higher than desirable, we suggest that managers adopt mitigation strategies, such as access control or abstraction of features deemed most risky (e.g., demographics).

In light of these observations, managers should note that it is difficult, if not impossible, to dictate a perfect recipe for designing a safe biorepository. However, diligence and pragmatism can help in designing an appropriate mix of technical and policy-based controls to mitigate identifiability risks.

Conclusions

We wish to stress that though only a limited number of privacy breaches have been reported, this does not imply that data is safe. It is difficult to detect when a re-identification has occurred and even more challenging to prove such an action in a legal setting. However, we can model a potential attacker's resources and knowledge and quantify risks. Using methods to assess the risk posed by those attack models we think most probable, we can tailor data sharing policies in light of real information.

Acknowledgments

This work was supported, in part, by grants 1R01LM009989 and 1U011HG004603 from the US National Institutes of Health.

References

Adam N, Wortman J. Security-control methods for statistical databases: a comparative study. ACM Comput Surv. 1989; 21:515–556.

[Accessed 27 May 2011] CODIS: the combined DNA index system. DNANews.org. 2011. http://dnanews.org/codis-the-combined-dna-index-system/

Bayardo, R.; Agrawal, R. Data privacy through optimal k-anonymity. Proceedings of the 21st IEEE International Conference on Data Engineering; 2005. p. 217-228.

Bellazi R, Zupan B. Predictive data mining in clinical medicine: current issues and guidelines. Int J Med Inform. 2008; 77:81–97. [PubMed: 17188928]

Benitez K, Malin B. Evaluating re-identification risk with respect to the HIPAA Privacy Rule. J Am Med Inform Assoc. 2010; 17:169–177. [PubMed: 20190059]

Benitez, K.; Loukides, G.; Malin, B. Beyond Safe Harbor: automatic discovery of health information de-identification policy alternatives. Proceedings of the ACM International Health Informatics Symposium; New York: ACM Press; 2010. p. 163-172.

Bexelius C, Hoeyer K, Lynöe N. Will forensic use of medical biobanks decrease public trust in healthcare services? Some empirical observations. Scand J Public Health. 2007; 35:442. [PubMed: 17786809]

Botkin J. Protecting the privacy of family members in survey and pedigree research. JAMA. 2001; 285:207–211. [PubMed: 11176815]

Burke W, Psaty B. Personalized medicine in the era of genomics. JAMA. 2007; 298:1682–1684. [PubMed: 17925520]

Cassa C, Schmidt B, Kohane I, Mandl K. My sister's keeper? Genomic research and the identifiability of siblings. BMC Med Genomics. 2008; 1:32. [PubMed: 18655711]

Chiang Y, Hsu T, Kuo S, Liau C, Wang D. Preserving confidentiality when sharing medical database with the Cellsecu system. Int J Med Inform. 2003; 71:17–23. [PubMed: 12909154]

Clayton D. On inferring presence of an individual in a mixture: a Bayesian approach. Biostatistics. 2010; 11:661–673. [PubMed: 20522729]

Clayton E, Smith M, Fullerton SM, et al. Confronting real time ethical, legal, and social issues in the Electronic Medical Records and Genomics (eMERGE) Consortium. Genet Med. 2010; 12:616–620. [PubMed: 20733502]

Collins F. Has the revolution arrived? Nature. 2010; 464:674–675. [PubMed: 20360716]

Currie P. Balancing privacy protections with efficient research: institutional review boards and the use of certificates of confidentiality. IRB. 2005; 27:7–12. [PubMed: 16425475]

Dankar, F.; El Emam, K. A method for evaluating marketer re-identification risk. Proceedings of the EDBT/ICDT Workshops; New York: ACM Press; 2010.

Eiseman, E.; Bloom, G.; Brower, J.; Clancy, N.; Olmstead, S. Case studies of existing human tissue repositories: "best practices" for a biospecimen resource for the genomic and proteomic era. Rand Corporation; Santa Monica: 2003.

El Emam K. Heuristics for de-identifying health data. IEEE Secur Priv Mag. 2008; 6:58–61.

El Emam K, Dankar K. Protecting privacy using *k*-anonymity. J Am Med Inform Assoc. 2008; 15:627–637. [PubMed: 18579830]

El Emam K, Jabbouri, Sams S, Drouet Y, Power M. Evaluating common de-identification heuristics for personal health information. J Med Internet Res. 2006; 8:e28. [PubMed: 17213047]

El Emam K, Dankar K, Issa R, et al. A globally optimal k-anonymity method for the de-identification of health data. J Am Med Inform Assoc. 2009; 16:670–680. [PubMed: 19567795]

Glaser J, Henley D, Downing D, Brinner K. Advancing personalized health care through health information technology: an update from the American Health Information Community's Personalized Health Care Workgroup. J Am Med Inform Assoc. 2008; 15:391–396. [PubMed: 18436899]

Golle, P. Revisiting the uniqueness of simple demographics in the US population. Proceedings of the ACM Workshop on Privacy in Electronic Society; New York: ACM Press; 2006. p. 77-80.

Green ED, Guyer MS. National Human Genome Research Institute . Charting a course for genomic medicine from base pairs to bedside. Nature. 2011; 470:204–213. [PubMed: 21307933]

Guttmacher A, Collins F. Realizing the promise of genomics in biomedical research. JAMA. 2005; 294:1399–1402. [PubMed: 16174701]

Haga S, O'Daniel J. Public perspectives regarding data sharing practices in genomics research. Public Health Genomics. 2011 (published online March 24). 10.1159/000324705

Hamburg M, Collins F. The path to personalized medicine. N Engl J Med. 2010; 363:301–304. [PubMed: 20551152]

Hansson S, Björkman B. Bioethics in Sweden. Camb Q Healthc Ethics. 2006; 15:285–293. [PubMed: 16862931]

Hindmarsh R, Abu-Bakar A. Balancing benefits of human genetic research against civic concerns: essentially Yours and beyond—the case of Australia. Pers Med. 2007; 4:497–505.

Homer N, Szelinger S, Redman M, et al. Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. PLoS Genet. 2008; 4:e1000167. [PubMed: 18769715]

Kaufman DJ, Murphy-Bollinger J, Scott J, Hudson KL. Public opinion about the importance of privacy in biobank research. Am J Hum Geneti. 2009; 85:643–654.

Kaye J. Police collection and access to DNA samples. Genomics Soc Policy. 2006; 2:16–72.

Kayser M, Schneider P. DNA-based prediction of human externally visible characteristics in forensics: motivations, scientific challenges, and ethical considerations. Forensic Sci Int Genet. 2009; 3:154–161. [PubMed: 19414162]

Kohane I, Altman R. Health information altruists—a potentially critical resource. N Engl J Med. 2005; 353:2074–2077. [PubMed: 16282184]

Kullo I, Fan J, Pathak J, Savova G, Ali Z, Chute C. Leveraging informatics for genetic studies: use of the electronic medical record to enable a genome-wide association study of peripheral arterial disease. J Am Med Inform Assoc. 2010; 17:568–574. [PubMed: 20819866]

Langella S, Hastings S, Oster S, et al. Sharing data and analytical resources securely in a biomedical research grid environment. J Am Med Inform Assoc. 2008; 15:33–373.

Lemke A, Wolf W, Hebert-Beirne J, Smith M. Public and bio-bank participant attitudes toward genetic research participation and data sharing. Public Health Genomics. 2010; 13:368–377. [PubMed: 20805700]

Lemrow S, Colditz G, Vaught J, Hartge P. Key elements of access policies for biorepositories associated with population science research. Cancer Epidemiol Biomarkers Prev. 2007; 16:1533–1535. [PubMed: 17684124]

Li G, Wang Y, Su X. Improvements on a privacy-protection algorithm for DNA sequences with generalization lattices. Comput Methods Programs Biomed. 201110.1016/j.cmpb.2011.02.013

Lin Z, Hewett M, Altman R. Using binning to maintain confidentiality of medical data. Proc AMIA Symp. 2002:454–458. [PubMed: 12463865]

Lin Z, Owen A, Altman R. Genetics: genomic research and human subject privacy. Science. 2004; 305:183. [PubMed: 15247459]

Lin Z, Altman R, Owen A. Confidentiality in genome research. Science. 2006; 313:441–442. [PubMed: 16873628]

Louie B, Mork P, Martin-Sanchez F, Halevy A, Tarczy-Hornoch P. Data integration and genomic medicine. J Biomed Inform. 2007; 4:5–16. [PubMed: 16574494]

Loukides G, Denny J, Malin B. The disclosure of diagnosis codes can breach research participants' privacy. J Am Med Inform Assoc. 2010a; 17:322–327. [PubMed: 20442151]

Loukides G, Gkoulalas-Divanis A, Malin B. Anonymization of electronic medical records for validating genome-wide association studies. Proc Natl Acad Sci USA. 2010b; 107:7898–7903. [PubMed: 20385806]

Lowrance W, Collins F. Ethics: identifiability in genomic research. Science. 2007; 317:600–602. [PubMed: 17673640]

Lunshof J, Chadwick R, Vorhaus D, Church G. From genetic privacy to open consent. Nature Rev Genet. 2008; 9:406–411. [PubMed: 18379574]

Mailman MD, Feolo M, Jin Y, et al. The NCBI dbGaP database of genotypes and phenotypes. Nat Genet. 2007; 39:1181–1186. [PubMed: 17898773]

Malin B. An evaluation of the current state of genomic data privacy protection technology and a roadmap for the future. J Am Med Inform Assoc. 2005a; 12:28–34. [PubMed: 15492030]

Malin B. Protecting genomic sequence anonymity with generalization lattices. Methods Inf Med. 2005b; 44:687–692. [PubMed: 16400377]

Malin B. A computational model to protect patient data from location-based re-identification. Artif Intell Med. 2007; 40:222–239.

Malin B. *K*-unlinkability: a privacy protection model for distributed data. Data Knowl Eng. 2008; 64:294–311.

Malin B, Sweeney L. How (not) to protect genomic data privacy in a distributed network: using trail re-identification to evaluate and design anonymity protection systems. J Biomed Inform. 2004; 37:179–192. [PubMed: 15196482]

Malin B, Karp D, Scheuermann R. Technical and policy approaches to balancing patient privacy and data sharing in clinical and translational research. J Investig Med. 2010; 58:11–18.

Malin B, Benitez K, Masys D. Never too old for anonymity: a statistical standard for demographic data sharing via the HIPAA Privacy Rule. J Am Med Inform Assoc. 2011; 18:3–10. [PubMed: 21169618]

McCartney C. Forensic DNA sampling and the England and Wales National DNA database: a sceptical approach. Crit Criminol. 2004; 12:157–178.

McCarty C, Chisholm R, Chute C, et al. The eMERGE Network: a consortium of biorepositories linked to electronic medical records data for conducting genomic studies. BMC Med Genomics. 2011; 4:13. [PubMed: 21269473]

McGuire A, Gibbs R. Genetics: no longer de-identified. Science. 2006; 312:370–371. [PubMed: 16627725]

McGuire A, Fisher R, Cusenza P, et al. Confidentiality, privacy, and security of genetic and genomic test information in electronic health records: points to consider. Genet Med. 2008a; 10:495–499. [PubMed: 18580687]

McGuire A, Hamilton J, Lunstroth R, McCullough L, Goldman A. DNA data sharing: research participants perspectives. Genet Med. 2008b; 10:46–53. [PubMed: 18197056]

Miler G. 2009 The looming crisis in human genetics. The Economist. Nov 13.

Miller E. Relative doubt: familial searches of DNA databases. Mich Law Rev. 2010; 109:291–348.

National Institutes of Health. NIH announces statement on certificates of confidentiality. 2002 Mar 15. NOT-OD-02-037.

National Institutes of Health. Final NIH statement on sharing research data. 2003 Feb 26. NOT-OD-03-032.

National Institutes of Health. Policy for sharing of data obtained in NIH supported or conducted genome-wide association studies (GWAS). 2007 Aug 28. NOT-O-07-088.

Ng P, Murray S, Levy S, Venter C. An agenda for personalized medicine. Nature. 2009; 461:724–726. [PubMed: 19812653]

Ollier W, Sprosen T, Peakman T. UK Biobank: from concept to reality. Pharmacogenomics. 2005; 6:639–646. [PubMed: 16143003]

Ossorio P. About face: forensic genetic testing for race and visible traits. J Law Med Ethics. 2006; 34:277–292. [PubMed: 16789949]

Phillips C, Salas A, Sanchez JJ, et al. Inferring ancestral origin using a single multiplex assay of ancestry-informative marker SNPs. Forensic Sci Int Genet. 2007; 1:273–280. [PubMed: 19083773]

Ritchie M, Denny J, Crawford D, et al. Robust replication of genotype–phenotype associations across multiple diseases in an electronic medical record. Am J Human Genet. 2010; 86:560–572. [PubMed: 20362271]

Roden D, Pulley J, Basford M, et al. Development of a large-scale de-identified DNA biobank to enable personalized medicine. Clin Pharmacol Ther. 2008; 84:362–369. [PubMed: 18500243]

Roses A. Pharmacogenetics and drug development: the path to safer and more effective drugs. Nat Rev Genet. 2004; 5:645–656. [PubMed: 15372086]

Samarati P. Protecting respondents identities in microdata release. IEEE Trans Knowl Data Eng. 2001; 13:1010–1027.

Sankararaman S, Obozinski G, Jordon M, Halperin E. Genomic privacy and limits of individual detection in a pool. Nat Genet. 2009; 41:965–967. [PubMed: 19701190]

Subcommittee on Disclosure Limitation Methodology, Federal Committee on Statistical Methodology. Statistical Policy Working Paper 22. Office of Management and Budget; 2005. Report on statistical disclosure limitation methodology. Revised by the Confidentiality and Data Access Committee

Sweeney L. Weaving technology and policy together to maintain confidentiality. J Law Med Ethics. 1997; 25:98–110. [PubMed: 11066504]

Sweeney L. *k*-anonymity: a model for protecting privacy. Int J Uncertain Fuzziness Knowl Based Syst. 2002a; 10:557–570.

Sweeney L. Achieving *k*-anonymity privacy protection using generalization and suppression. Int J Uncertain, Fuzziness Knowl Based Syst. 2002b; 10:571–588.

U.S. Department of Health and Human Services. Standards for privacy of individually identifiable health information, final rule. Federal Register. 2002; 45 CFR:160–164.

Vinterbo S, Ohno-Machado L, Dreiseitl S. Hiding information by cell suppression. Proc AMIA Symp. 2001:26–730.

Wang D, Liau C, Hsu T. Medical privacy protection based on granular computing. Artif Intell Med. 2004; 32:137–149. [PubMed: 15364097]

Wang, R.; Li, Y.; Wang, X.; Tang, H.; Zhou, X. Learning your identity and disease from research papers: information leaks in genome wide association study. Proceedings of the ACM Conference on Computer and Communications Security; New York: ACM Press; 2009. p. 34-55.

Willenborg, L.; De Waal, T. Springer Lecture Notes in Statistics. Springer; New York: 1996. Statistical disclosure control in practice.

Wolf L, Zandecki J. Sleeping better at night: investigators' experiences with certificates of confidentiality. IRB. 2006; 28:1–7. [PubMed: 17849658]

Zerhouni E, Nabel E. Protecting aggregate genomic data. Science. 2008; 322:44. [PubMed: 18772394]

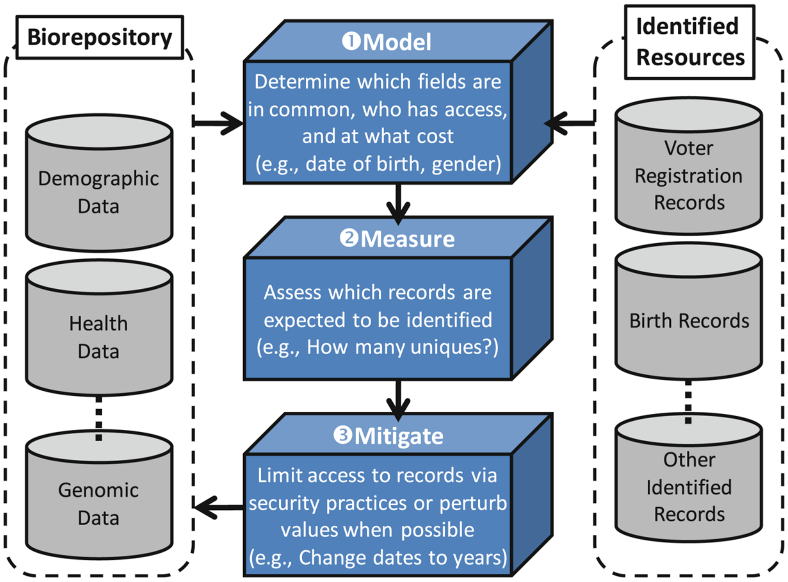
Biorepository
Demographic Data
Health Data
Genomic Data
❶Model
Determine which fields are in common, who has access, and at what cost (e.g., date of birth, gender)
❷Measure
Assess which records are expected to be identified (e.g., How many uniques?)
❸Mitigate
Limit access to records via security practices or perturb values when possible (e.g., Change dates to years)
Identified Resources
Voter Registration Records
Birth Records
Other Identified Records

Fig. 1.

A general framework assessing and mitigating health data re-identification risks. Note that mitigation is performed with respect to the information in the biorepository. This is because data that are a NSl ready public can not be controlled by biorepository managers

Table 1

Principles to assist experts in the determination of the identifiability of health information

Principle	Description	Examples
Replication	Prioritize health information features into levels of risk according to the chance it will consistently occur in relation to the individual	Low: results of a patient's blood glucose level test will vary High: Demographics of a patient (e.g. birthdate) are relatively static
Resource availability	Determine which external resources contain the patients' identifiers and the replicable features in the health information, as well as who is permitted access to these resources	Low: The results of laboratory reports are not often disclosed with identity beyond healthcare environments High: Patient identity and demographics are often in public resources, such as vital records—birth, death, and marriage registries.
Distinguishability	Determine the extent to which the subject's data can be distinguished if health data is disseminated	Low: It has been estimated that the combination of *Year of Birth, Gender,* and *3-Digit ZIP Code* is unique for approximately 0.04% of residents in the United States (Sweeney 2007). This means that very few residents could be indentified through this combination of data alone High: It has been estimated that the combination of a patient's *Date of Birth, Gender,* and *5-Digit ZIP CODE* is unique for over 50% of residents in the United States (Golle, 2006, Sweeney 2002a, b). This means that over half of US residents could be uniquely described just with these three data elements

Table 2

Technical approaches to anonymize demographic and clinical information supplied to biorepositories

Approach	Privacy principle and data on which applied	Transformation method
El Emam and Dankar (2008)	*k*-anonymity and approximate *k*-map applied to demographics	Generalization and suppression
Benitez et al. (2010); Dankar and El Emam (2010)	Limit average re-identification risk when sharing demographics	Generalization
Loukides et al. (2010b)	*k*-based model applied to potentially distinguishing diagnosis codes	Generalization and suppression