Technical and Policy Approaches to Balancing Patient Privacy and Data Sharing in Clinical and Translational Research

Bradley Malin, PhD, David Karp, MD, PhD,† and Richard H. Scheuermann, PhD‡*

Introduction: Clinical researchers need to share data to support scientific validation and information reuse and to comply with a host of regulations and directives from funders. Various organizations are constructing informatics resources in the form of centralized databases to ensure reuse of data derived from sponsored research. The widespread use of such open databases is contingent on the protection of patient privacy.

Methods: We review privacy-related problems associated with data sharing for clinical research from technical and policy perspectives. We investigate existing policies for secondary data sharing and privacy requirements in the context of data derived from research and clinical settings. In particular, we focus on policies specified by the US National Institutes of Health and the Health Insurance Portability and Accountability Act and touch on how these policies are related to current and future use of data stored in public database archives. We address aspects of data privacy and identifiability from a technical, although approachable, perspective and summarize how biomedical databanks can be exploited and seemingly anonymous records can be reidentified using various resources without hacking into secure computer systems.

Results: We highlight which clinical and translational data features, specified in emerging research models, are potentially vulnerable or exploitable. In the process, we recount a recent privacy-related concern associated with the publication of aggregate statistics from pooled genome-wide association studies that have had a significant impact on the data sharing policies of National Institutes of Health-sponsored databanks.

Conclusion: Based on our analysis and observations we provide a list of recommendations that cover various technical, legal, and policy mechanisms that open clinical databases can adopt to strengthen data privacy protection as they move toward wider deployment and adoption.

Key Words: clinical research, translational research, databases, privacy

(*J Investig Med* 2010;58: 11–18)

A number of organizations, distributed around the globe, have invested considerable effort to construct information technology infrastructure to support the management and analysis of data on human participants enrolled in clinical and translational research studies.[1] Organizations are now moving toward models of broader data sharing and accessibility through open-access translational research information systems (OTRISs). Open-access translational research information systems are dynamic and evolving in technical implementation and oversight but have a common goal of establishing data warehousing infrastructure to facilitate the rapid dissemination of research findings. They aim to integrate a variety of data types, such as experimental information derived from laboratory experimentation (eg, genome sequence, gene expression, and proteomics data) with rich clinical phenotypes. Open-access translational research information systems further aim to integrate data from various laboratories and other resources so that the research community has access to a broad range of datasets to validate and reanalyze published findings, as well as mine for novel clinically relevant discoveries. Thus, it is the intention of OTRIS managers to make their systems and, to the extent to which it is possible, the data within freely accessible as a resource to the public.

Open-access translational research information systems raise complex ethical, legal, and social issues that developers, managers, and scientists associated with these systems will need to consider as software engineering and scientific investigation move forward.[1–3] Recent meetings have solicited information from ethicists, informaticists, lawyers, and biomedical scientists to characterize various issues associated with the construction of database archives ranging from informed consent to attribution of property to the identifiability of human participants in supported research projects.[4] In this paper, we elaborate on the data privacy issues in the context of OTRISs. We recognize that a complete solution will require further investigation on ethical, social, and legal components of the problem, but we use this forum to illustrate how policy and technology can be combined to resolve data sharing and privacy goals.

It has been stressed that the availability of OTRISs for widespread use is contingent on the protection of patient anonymity.[5] Although biomedical privacy policies and technologies exist, various studies suggest they are ill-equipped for environments that centralize detailed patient-specific data.[6] Moreover, recent forensics science research[7,8] has prompted significant changes to data sharing policies for various OTRISs, most notably the database of Genotype and Phenotype (dbGaP)[9] at the US National Library of Medicine.[10] In the face of such threats, one must question if there are potential privacy vulnerabilities for other emerging resources. Furthermore, if such threats do exist, then what are the measures, from both technical and policy perspectives, that should be explored to mitigate them?

In this paper, we illustrate how OTRISs are vulnerable, but it is important to note that not all emerging OTRISs are susceptible to privacy violations in the same manner. In addition, the power that responsible policies and oversight can provide in mitigating threats that remain in de-identified research settings should not be neglected. The issues raised and potential solutions offered in this paper are applicable to many informatics resources intending to share clinical and biological data for translational

From the *Department of Biomedical Informatics, School of Medicine, Vanderbilt University, Nashville, TN; †Division of Rheumatology, Department of Internal Medicine, and ‡Division of Biomedical Informatics, Department of Pathology, University of Texas Southwestern Medical Center, Dallas, TX.

Received September 28, 2009, and in revised form November 4, 2009.
Accepted for publication November 6, 2009.

Reprints: Bradley Malin, PhD, Department of Biomedical Informatics, School of Medicine, Vanderbilt University, Suite 600, 2525 W End, Nashville, TN 37203. E-mail: b.malin@vanderbilt.edu.

Supported by the following grants from the US National Institutes of Health: N01AI40076, R01LM009989, U01HG004603, UL1RR023468, and UL1RR024982.

ISSN: 1081-5589
DOI: 10.231/JIM.0b013e3181c9b2ea

research purposes, and, where possible, we draw on examples from emerging OTRISs to demonstrate their potential application.

POLICIES AND REQUIREMENTS

Before we address technical issues, it is important to note the regulatory landscape. Data collected, shared, and used within OTRIS will be subject to various regulatory controls. The appropriateness of such controls depends upon from where the data will be derived. In particular, there are several primary privacy and data sharing policies that OTRIS managers must be cognizant of as they move forward. The following is an introduction to some of the relevant regulatory issues at play and should not be considered a comprehensive list.

NIH Data Sharing Policy

The National Institutes of Health (NIH) Data Sharing Policy was designed to increase access to data collected through, or studied with, federal funding.[11] The policy applies to all projects that receive at least $500,000 in annual direct funding. According to the policy, data must be shared in a de-identified format in a manner similar to the Safe Harbor model as defined in the Privacy Rule of the Health Insurance Portability and Accountability Act (HIPAA; discussed later). The data sharer must also remove information for which there is prior knowledge that it could be used to determine the identity of the subjects. Some investigators have argued that the sensitivity of their data sets and the lack of ability to provide provable privacy guarantees are sufficient to opt out of data sharing.

NIH GWAS Policy

Genomewide genetic scans of sequence variations have become important, but costly, research tools for the biomedical community. The NIH created a specific policy for the collection and sharing of data derived from, or studied in, genomewide association studies (GWAS).[12] Similar to the 2003 Data Sharing Policy, the GWAS policy was defined such that it applies to any project regardless of funding level in which genomewide genetic scans are produced or studied. The NIH has since designated the dbGaP as the repository to which NIH-sponsored investigators should submit their GWAS records. As in the NIH Data Sharing Policy, GWAS data must be de-identified before dissemination.

The NIH has recognized that genomic data itself may lead to the re-identification of an individual. Thus, users of GWAS data sets in the dbGaP must sign a contractual use agreement that explicitly prohibits nonsanctioned uses and attempts to identify subjects (discussed later). Other NIH groups and repositories are applying similar use agreements to assign legal constraints to the use of information stored in their OTRIS.

HIPAA Privacy Rule and the Many Forms of Data Sharing

In the United States, when a covered entity, as defined by HIPAA (eg, healthcare providers, health data clearinghouses, and other groups), wishes to share data collected in the context of clinical activities, it must adhere to the Privacy Rule.[13] The regulation outlines several routes by which personal health information can be shared without patient consent for secondary research purposes: (1) safe harbor, (2) limited data set, and (3) statistical certification.

The Safe Harbor standard allows covered entities to publicly share data once it is stripped of an enumerated list of 18 types of personal identifiers. These include explicit identifiers (eg, names), quasi-identifiers (eg, dates and geocodes), and traceable elements (eg, medical record numbers). Neither clinical nor genomic data are explicitly labeled as a personal identifier, and it has been debated if such data can be released under this policy.[14] For years, clinical data have been shared in public resources such as hospital discharge databases.[15,16] Similarly, person-specific DNA sequences have been disclosed to public repositories, such as those at the National Center for Biotechnology Information.[17]

Various groups argue against disclosing data via Safe Harbor based on the observations that the usefulness of such data for certain types of studies (eg, epidemiology) is questionable but also out of re-identification concerns.[5,18] Rather, an alternative called the Limited Data Set standard is advocated, which allows covered entities to share more detailed data, including dates and zip codes. The tradeoff is that data recipients must enter into an acceptable use contract that prohibits re-identification. Although this policy is appropriate for trusted investigators, as the quantity of data and number of investigators granted access increases, such an approach may become infeasible to manage. Moreover, this policy neither prevents a recipient from attempting re-identification nor assesses the risk of re-identification.

The Statistical Standard allows sharing data in any format, provided an expert certifies that "the risk is very small that the information could be used by the recipient, alone or in combination with other reasonably available information, to identify an individual."[13] Methods to quantify risks have been researched,[7,18] but no standards have emerged. One disclosure control method that has been considered is to perturb DNA sequences, for example, AACCTATA shared as AATCAATA.[19] The intuition is that as the quantity of perturbation increases, the likelihood that an investigator can determine the original sequence decreases, implying greater privacy protection. The tradeoff, however, is that perturbation can potentially obscure, or worse, lead to false associations. Thus, it could diminish the utility and scientific credibility of the resource. A second criticism of such a protection approach is that research has shown that certain types of perturbation can be filtered to reliably infer the original data.[20] Despite such problems, data protection based on scientific models can be achieved, but care must be taken to design them with formal principles.

RE-IDENTIFICATION MODELS, METHODS, AND APPLICATIONS

As we alluded to, data that are de-identified according to the aforementioned polices can be re-identified to the individuals from which the data were derived via numerous routes. As we illustrate in Figure 1, re-identification is a process and requires the

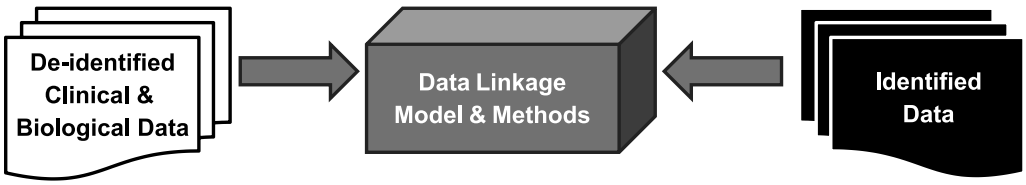

FIGURE 1. General model of data re-identification. There are 3 conditions that need to be satisfied: the ability to distinguish an individual's record in (1) de-identified and (2) identified resources, and (3) a mechanism for relating (or linking) data from the resources.

satisfaction of certain conditions. First, it requires that the de-identified data are unique or "distinguishing." In other words, we must be able to pinpoint an individual in a group of size *n* people or less. Genomic sequence data, for instance, and possibly other laboratory and molecular expression data, are often highly distinguishing. However, it needs to be recognized that the ability to distinguish data is, by itself, insufficient to claim that the corresponding individual's privacy will actually be compromised. This is because of the second condition, which is that we need a naming resource. Without such a resource, there is no way to link the de-identified data to an identity. *Finally, for the third condition, we need a mechanism to relate the de-identified and identified resources. Inability to design such a relational mechanism would hamper an adversary's opportunity to achieve success to no better than random assignment of de-identified data and named individuals.

There are many situations in which de-identified biomedical information can be re-identified to the patient from whom it was derived without hacking or breaking into private health information systems. For instance, in the mid-1990s, it was shown that de-identified hospital discharge records, which were publicly available at the state level, could be linked to identified public records in the form of voter registration lists. The result received notoriety because it led to the re-identification of the medical status of the governor of the Commonwealth of Massachusetts.[21] This attack was achieved by linking the resources on the seemingly innocuous, but common, fields of a patient's date of birth, sex, and zip code. Various estimates indicate that the uniqueness of this combination of attributes in the US population is somewhere between 65% and 87% and even more unique for certain subpopulations.[22,23]

Risk of Identification

One of the responses to the discharge record attack was the HIPAA Safe Harbor policy. However, it should be recognized that even the suppression of all enumerated features fails to prevent all re-identifications. In many instances, there are residual features, including the remaining demographics (eg, race, year of birth, state of residence, and sex) that can lead to identification. However, the extent to which residual features can be applied to re-identification is context dependent and relies on the availability of the fields that can be leveraged in the attack. In Table 1, we provide some general guidelines to consider when assessing the re-identification risk of data in OTRIS. In general, it helps to partition the person-specific features into classes of relatively "high" and relatively "low" risks. We recognize that risk is more of a continuous variable, but this type of dichotomization helps illustrate how context impacts risk. Beyond riskiness of attributes, it is important to understand the routes by which data can be linked to naming sources or sensitive knowledge can be inferred, as we review below.

What is a High-Risk Identifier?

Higher-risk features are those that are documented in multiple environments and are publicly available. These are features that can be exploited by any recipient of such records. For instance, patient or research subject demographics are high-risk identifiers. Even the demographics that are permissive under the Safe Harbor policy leave certain individuals in a unique status and thus at nontrivial risk for identification through public resources that contain similar features, such as birth, death, marriage, voter, property assessor records, and more.

What is a Low-Risk Identifier?

Lower-risk features are those that do not appear in public records and are less readily available. For instance, clinical features, such as an individual's diagnoses and treatments are relatively static (ie, because they are often mapped to standard coding terminologies for billing purposes), and can manifest in de-identified resources, such as the aforementioned hospital discharge databases, and in identified resources, such as electronic medical records. Combinations of diagnosis and treatment codes, or temporal dependencies, can uniquely characterize a patient in a population,[24] but the combination of large quantities of standard code (ie, >5 codes) tied to identified records is available to a much smaller group of individuals than the general public. Moreover, this select group of individuals may be relatively more trustworthy, such as care providers and business associates of the organization that generated the documented features. Additional disincentives may exist as well, such as HIPAA-related penalties that are applied in the event an individual willingly violates the terms of employment to commit a breach of privacy.

Where Do Data Derived From Biosamples Come Into Play?

When OTRISs include data derived from biological samples, the situation becomes a bit more complex. In certain instances, the information that is associated with genomic and expression data, particularly genomic data derived from a clinical setting, permits relationships to be established between de-identified and identifiable resources. Yet, it should be recognized that this is not always the case. The following is a summary of several attacks, with further details available elsewhere.[6]

Genotype-Phenotype

There exists an inherent relationship between certain genomic data sequences and physical phenotypic manifestations. A clinical phenotype may be described in biomedical coding standards such as the *International Classification of Diseases* and may be disclosed in various settings, including semiprivate data such as administrative or insurance records as well as more public records such as hospital discharge databases.

Familial Information

A second type of attack is made possible because genomic data are increasingly disseminated in the context of familial information. This practice is common in gene hunting expeditions. Familial information could be represented in the format of a de-identified pedigree, which reports sex, disease status, and the death status of the family members. At the same time, there is a variety of publicly available identified information available. One particular resource that has been exploited for identifiers is obituaries, which have a wide coverage on a population and are often free to post in online, searchable newspapers. Such resources tend to include information on the recently deceased individual as well as family relations.[25]

Trails and Location-Based Patterns

Many patients (and research participants) are transient and visit multiple institutions providing care. As such, a patient's location visit pattern is often distinguishing and facilitates what has been termed a "trails" attack.[26] In this scenario, a patient visits multiple hospitals, where his clinical and DNA-related data are

*We recognize that the lack of a readily available naming resource does not imply that data are sufficiently protected from re-identification. Nonetheless, it does indicate that it is much harder to identify an individual, or group of individuals, given the resources at hand.

TABLE 1. Summary of OTRIS Data Re-Identification Assessment Mechanisms

Replication	Prioritize OTRIS data attributes into different levels of risk according to their replicability (eg, molecular expression data are less replicable than genomic sequence data).
Resources	Determine which external resources contain the subject's identities and the replicable attributes in the OTRIS data.
Distinguish	Determine the extent to which the subjects' de-identified data can be distinguished in the OTRIS (eg, it is easier to distinguish an individual using 100 SNPs as opposed to one using 1 SNP)
Access	Determine who has access to the identified resources. Are the data publicly available? Is it a more private resource?
Assess risk	The greater the replicability, the availability, and the distinguishability of OTRIS data, the greater the risk for re-identification (and vice versa).

collected. The facilities forward the de-identified DNA records, tagged with the submitting institution, to a public centralized data bank.[27–29] In addition, the hospitals send identifiable discharge records, including patient demographics and diagnoses, to a discharge database.[30] Even if there is no clear biomedical relationship between the diagnosis codes and sequence markers in the DNA, we can track the hospitals a patient has visited (ie, the trail) in the discharge data and the DNA records in the repository.[26] Notably, this attack is generalizable in that trails can manifest in a number of environments.[31]

Genome Sequence Data

Genome sequence data are increasingly applied in clinical research. However, it is also a well-known distinguishing feature unto itself. Lin et al.[19] demonstrated that only a small number (less than 100) of single nucleotide polymorphisms (SNPs) is required to uniquely characterize an individual in the entire world's population. Single nucleotide polymorphism data are increasingly found in ancestry, clinical and molecular phenotype, and pharmaceutical efficacy association studies. Thus, if an adversary has access to an identified DNA sequence, it may be possible to learn additional information about the individual from the de-identified data in the association studies.

In recognition of this fact, the dbGaP decided to publicly disseminate SNP-clinical status correlations for various data sets only as aggregated results. Specifically, for each data set, and for each individual SNP, they publicly posted the proportion of the population that was diagnosed with (or without) a clinical feature (eg, immunodeficiency disorder) and the corresponding SNP value.

However, as was recently demonstrated by Homer et al.[7] and Jacobs et al.,[8] such an approach does not prevent privacy threats. They demonstrated what we call a "pool attack," where the information on several thousand SNPs could be used to determine if an identified individual's DNA was in the set of clinically positive cases, the set of clinically negative cases, or neither of the above. Moreover, the approach involved in the attack is applicable to any environment in which aggregate statistics on biomedical data sets are available. Details of their attack are beyond the scope of this paper but can be found elsewhere. This attack is important to note because it had significant impact on the dbGaP's public data access policy (as described in the following section).

Laboratory Reports and Expression Data

The previous types of data and attacks may compromise privacy because they are sufficiently replicable and available in multiple data sets. However, data stored in many OTRISs is also expected to consist of functional genomics data (eg, gene expression microarray data) derived from laboratory testing. Although such data may be unique and located in multiple data sets, the extent to which these data are replicable is questionable. To the best of our knowledge, there is limited research that addresses the precision of repeated functional genomics tests. However, if such information is not adequately replicable, these data may be considered less risky to share than sequence data.

OBSERVATIONS

Before proposing specific recommendations regarding technologies and policies to improve data privacy protections, we wish to return to the pool attack and highlight several results and policy decisions. The pool attack did not involve a compromise of identity because the adversary was already in the possession of the subject's identity and genomic data. However, the attack resulted in a breach of confidentiality because the subject did not inform the adversary of their clinical status. Given the reported accuracy of the attack, the NIH felt they could not publish statistical summaries of SNP-clinical class correlations without violating the privacy principles stated in their data sharing policies. As a consequence, the NIH removed all summary statistics from the public version of the dbGaP.[10] Following the lead of the NIH, the Wellcome Trust, the main biobanking and human genomic data dissemination agency in the UK, followed suit. The policy changes received significant attention from the popular media.[32–35] Although privacy advocates have lauded these actions, there are several reasons why this response is not necessarily appropriate for every OTRIS.

Observation 1: The Attack Requires an Identified Reference Sample

The attack is a feasible one in that it can be achieved given relatively open data sharing strategies. In fact, the approach is generalizable to other types of information derived from biological samples. However, the question remains as to what the likelihood of such an attack is given today's climate. To achieve the attack, the adversary needs access to an identified DNA sequence, which begs the question of who would be in possession of such information? It has been suggested that such information could be available through forensic investigations, but it is unclear if forensic specialists would be sufficiently motivated to learn clinical information about the subject in question. Second, it has been noted that individuals beyond the forensic realm could collect biological samples, subsequently sequence and use the resulting information, but the economic and computational barriers are nontrivial, and it is not clear that anyone would attempt to mount such an attack. This is not to say that such an attack could not be committed by motivated individuals but that the context for executing such an attack has yet to be clearly voiced. We recognize that biological data will be increasingly available as high-throughput genomic sequencing technology becomes cheaper and more mainstream, but at the present time, the threat is believed to be more theoretical than practical.

Observation 2: Regulating Aggregate Results and Microdata in the Same Manner Is a Potentially Restrictive Administrative Model

Investigators conducting NIH-sponsored GWAS are encouraged to submit their de-identified records to the dbGaP.[12] The

result is that the dbGaP stores data sets for a number of NIH institutes. Initially, the dbGaP defined its access policy according to a 2-tier model. The first tier consisted of public information, which included summary information for each data set, including data collection mechanisms, the types of demographic, clinical, and biological information collected, and summary statistics for the various classes of individuals. This was designated as public information that was readily available on the dbGaP website. The second tier of access was for person-level records, or "microdata." To access this information, investigators must proceed through a formal evaluation process. The process begins when a new investigator submits a request to access the records in a data set. The application is sent to an NIH data access committee (DAC). Because each data set may have unique use limitations and may have been sponsored by a different NIH institute, the investigator may need to make multiple requests for multiple datasets.

When the NIH decided that summary statistics for data deposited in the dbGaP would no longer be accessible through the first, or public, access level, such information was moved to the second tier. However, the DAC model was designed to handle requests for individual-level data sets to validate or explore specific hypotheses, not requests to mine for new knowledge across data sets that are unrelated in the initial reasons for their collection. Thus, this approach to managing summary information could limit large-scale data mining and hypothesis generation–driven research methodologies that are gaining popularity in the biomedical domain.

Observation 3: Technical and Statistical Measures Can Be Applied to Disseminate Person-Specific GWAS Data With Privacy Risk Guarantees

From a technical perspective, the removal of summary statistics from the public realm created an "all or nothing" data access setting. Initially, researchers were permitted access to all of the SNPs and the relative occurrence of variant statistics, but after the policy change, researchers were shuttled into a "nothing" model, in which no statistical information on any SNP could be reported. Given the current manner in which data protection is achieved and the existing protection technologies on the market, this is a logical situation. However, as recent research suggests, there is room to create a gray solution that resides within this policy space through the use of risk analysis strategies. Consider that, as we mentioned earlier, if an adversary has access to summary information about a single SNP, then the likelihood the adversary can map an identified DNA sequence to the affected, nonaffected, or none of the above classes is significantly hampered (Table 2). If provided with summary information about 2 SNPs, the probability that the adversary could link the identified record to one of the classes would be greater but still extremely small. As we increase the number of SNPs that an investigator is permitted to have access to, the probability of linkage will increase. If data managers could determine the level of risk they are willing to tolerate, then they could disseminate information on a subset of SNPs consistent with their level of risk tolerance. This is precisely the basis for a formal and provable data protection strategy in accordance with the statistical data protection standard mentioned in the previous section.

TABLE 2. Contingency Table Reporting the Counts of Variant Frequency per Phenotype Class for SNP *i*

		Phenotype Class		
		Affected	**Nonaffected**	**Margins**
SNP position *i*	α (eg, A)	a_i	b_i	$a_i + b_i$
	β (eg, T)	c_i	d_i	$c_i + d_i$
	Margins	$a_i + c_i$	$b_i + d_i$	N

TABLE 3. Summary of Technical and Policy Approaches for OTRIS Data Privacy Protection

1	Publish aggregate statistics of biomedical data only when there is low risk of exploiting the data for linkage
2	Assess the replication reliability of molecular data
3	Define access policies and assess credentials of users
4	Define use agreements
5	Solicit informed consent for future data use when appropriate
6	Formalize liability and redress procedures
7	Establish auditing practices
8	Use multiple levels of data detail and oversight when possible
9	Adopt technically formal re-identification risk mitigation approaches (eg, *k*-anonymity)

RECOMMENDATIONS AND FUTURE DIRECTIONS

The earlier sections provided a high-level analysis of the existing threats and potential opportunities for OTRIS. This section formalizes specific recommendations regarding technologies and policies to improve data privacy protections. There is no single solution that will address all privacy and identifiability issues, but a combination of technical, policy, and legal mechanisms will help ameliorate potential problems.

As biomedical data sharing increases and systems move toward open access, there are certain guidelines and recommendations we believe OTRIS should consider. The following recommendations are briefly summarized in Table 3.

Publishing Aggregate Statistical Information for Known Replicable Features Only When the Risk of Exploiting Such Features Is Sufficiently Low

Given current NIH policies, it is recommended that OTRIS not post pooled statistical information on publicly accessible web servers regarding static replicable features that are easy to derive from biological information, such as genomewide SNP scans. Although the risk of an individual actually applying such information in a linkage attack is unknown, the posting of such information will be in direct contradiction of policies adopted by similar NIH repositories and recent statements of the NIH director.†

Assess the Replicability of Molecular Data Types

As noted earlier, functional genomics data may be the focus of a given database repository. It is anticipated that the reliability of data replication will be data type specific, and it is thus recommended that OTRIS management discuss this issue with the scientists submitting the data. If the data are unreliably replicable, then the risk of publishing such information is less of a concern and the OTRIS may justify less strict oversight to access the data.

†There are emerging algorithms, implemented in working software, that can help data managers determine which features can be disclosed while ensuring that the probability of classifying an individual is below a predefined threshold.[36] Yet, until such approaches are tied to appropriate policy models, their implementation will be limited.

If, on the other hand, such data are reliably replicable, then the data in the OTRIS could be subject to a pool attack. In this case, it is recommended that such data should not be shared publicly.

Establish Polices for Assessing Credentials of Data Users and Committees to Institute the Policies

It is recommended that formal data access policies be established and published on the appropriate OTRIS management's website or made available through the appropriate regulatory bodies. In association with a formalized policy, it is further recommended that OTRIS establish a DAC that reviews applications for access to data. This committee may be designed in a similar manner to the dbGaP DAC but should be tailored to the needs of the repository. Individuals that serve on this committee could be drawn, to the extent it is possible, from the following classes:

a. Ethicists
b. Lawyers/Counselors
c. Scientists that deposited data into the OTRIS
d. Program managers from funding agencies, including
   i. Scientific research officials
   ii. Science policy officials

Additional groups that may be represented on such a committee could consist of

a. Patients/Community advocates for whom data in the OTRIS correspond;
b. External advisors from related biomedical repositories (eg, the dbGaP) or projects (eg, the GAIN network,[37] or the EMERGE consortium[38]).

If the resource determines that data should be made available to anyone with a legitimate request, the access committee's role may only need to define what such requests correspond to and perform expedited reviews of requests for data access.

Define Use Agreements

It is recommended that the OTRISs determine what is considered acceptable use with respect to accessed data. Such information should be codified and explicitly defined in a data use contract that is agreed upon by the data recipient. It is recommended that the OTRISs work with legal experts with experience in this area to establish appropriate terms.

Informed Consent Should Describe, to the Extent Possible, the Risks of Data Aggregation and Reuse

De-identification and controlled access are essential aspects of legal and ethical data reuse from existing research databases and electronic health records. At the same time, however, much information will come from prospective mechanistic and translational studies, including GWAS. In these cases, the informed consent process must disclose the potential data sharing mechanisms described in this paper. It should be recognized that it may not be possible to describe all of the future users of a subject's de-identified data, and it may not be legally possible for subjects to consent to unspecified future uses.[39] However, subjects should be entitled with the opportunity to authorize future uses of their data for particular types of studies and withhold permission for others.[40] Documentation of such understanding by subjects when they enter into research studies will assist institutional review boards and ethics committees to provide the necessary certification when data are shared according to NIH policies. Moreover, clear demonstration that subjects in genetic research know and understand the potential of re-identification may lessen the regulation imposed in response to various privacy-invading mechanisms, such as the aforementioned pool attack.

Formalize Liability Requirements and Procedures for Redress

Although data shared through the OTRIS may be de-identified, it may be potentially re-identifiable. As such, the resource needs to determine the extent to which it is willing to assume liability for misuse of data. For instance, if a data recipient actually performs re-identification of a record, and such a re-identification becomes known, there should be a standing policy for how best to address and/or reprimand the user. Responding to the situation may be handled by the OTRIS itself, the recipient's home institution (if one exists), the originating institution of the data, or by any combination of the parties. Regardless, policies and procedures need to be established and agreed upon by all involved. Again, the resource should work with the appropriate legal specialists and stakeholders in this activity.

Establish Auditing Practices

Even if the OTRIS chooses to make data public (or semipublic), it should enable auditing capability. In doing so, the OTRIS should assign unique login and passwords for each data user and log their activities in immutable audit logs. The resource should also determine when and how to audit users of the OTRIS. In most cases, data users will not act maliciously, but they may violate terms of service or best practices of use without realizing it.

Consider Multiple Levels of Accessibility and Oversight to Access Information

It is recommended that the managers of the OTRIS determine what they consider to be acceptable levels of risk and realistic vulnerabilities to the data in the system (the examples of high- and low-risk identifiers discussed earlier can help guide this discussion). If possible, the OTRIS may wish to provide access to different levels of data detail. For instance, it could provide access to aggregate statistics at one level and detailed microdata at another. At all levels, the aforementioned access committee should be involved. Moreover, it should be noted that managing aggregate statistical features of biological or molecular data in the same manner as the actual microdata is a potentially overprotective research-limiting step. For resources of lower risk (such as aggregates), the committee may choose to apply expedited reviews to ensure that the requests are in line with acceptable use policies, similar to those applied by institutional review boards. The goal is to minimize the amount of time that the committee needs to spend reviewing an individual's request to access data. In contrast, for access to more detailed information, the committee may use a more stringent review process and require additional restrictions on data access and transferability.

It should be recognized that there is no universal solution to mitigate identifiability. There is no definite set of data attributes that, if suppressed, will guarantee protection from the data being re-identified. Rather, it is recommended that risk estimates be performed to determine the level of risk involved with sharing the data (note: this risk is data dependent and not attribute dependent), and these risks should be deemed acceptable to the data managers, whether it is the investigators sharing data to the OTRIS, the sponsoring agency's program managers, or OTRIS administrators.

Adopt Technical Approaches to Mitigate Re-Identification Risk

As mentioned earlier, different types of data lead to different linkage concerns. Some data can be linked to publicly available data, especially demographics. It should be recognized that even if data shared via a database resource adhere to HIPAA's Safe Harbor

levels of protection, there is no guarantee the data are impregnable to re-identification. Thus, if there is a concern that someone would attempt to discredit the OTRISs by identifying a single record in the database, then managers should consider disclosing data according to a more formal data protection model. One manner by which the OTRISs can formally mitigate risk is to generalize and/or suppress data to ensure that each record corresponds to a certain number of people (ie, a minimum bin size).

Technically, the resource may consider a formal protection model such as *k*-anonymity.[41] In this model, the data protector chooses a "*k*" that specifies the risk deemed to be acceptable; specifically, *k* corresponds to how many people data managers want each specific record to link to. The question remains, however, as to the appropriate level *k* to be chosen. For some guidance, the various statistical agencies have suggested that approximately 5 seems to be an acceptable solution. Whether or not this is directly applicable to life sciences data remains an open question. If deemed acceptable, this is a solution that can be tailored to any data set in an OTRIS. In other words, *k* could be made dependent on the sensitivity of the data in question or the amount of harm that could be committed through the data.

The benefits of a privacy model, such as *k*-anonymity, are that it (1) naturally relates to the HIPAA protection policy of the statistic or scientific standard and (2) requires the data holder to cognitively be involved in the protection of data. The drawbacks are that (*a*) it is not clear how *k*-anonymization affects the utility of the data for translational hypothesis generation (and data mining in particular) and (*b*) it is not clear how *k*-anonymized data can be analyzed with typical statistical packages software for complex data types. Nonetheless, many biomedical research or application, of common genetic and clinical variants such that a formal data protection model may sufficiently preserve enough biomedical information for future investigators. Additional research is necessary to determine when this technical method of data protection is applicable.

The policy and technology recommendations we have outlined can be combined for flexible control. The recommendations should be used as the OTRIS deems necessary. The main goal is too strike an appropriate balance where the technical aspects of data protection are complemented with acceptable use and oversight policies. If data users are more trusted, then data may be disseminated in a more specific form with stringent use contracts and if users are deemed to be less trusted, then data may be disclosed in a more aggregated form with weaker use contracts.

CONCLUSIONS

In conclusion, there is increasing awareness by all of the various stakeholders involved in human studies research—research sponsors, investigators, and subject participants—that to maximize the return on the investment that all parties make in clinical research, it is advantageous to make the results widely available to the research community. A variety of OTRIS resources have been established to facilitate the sharing and reuse of these valuable data. However, whereas the benefits of data sharing are recognized, the requirements for maintaining the autonomy, privacy, and confidentiality of the research participants must also be addressed. We have presented a series of recommendations regarding both technical and policy approaches designed to minimize the risk of participant re-identification from clinical research data. By adopting these recommendations, OTRIS can balance the benefits gained by data sharing while minimizing the risk to the research participants.

ACKNOWLEDGMENTS

The authors thank Dr. Ellen Wright Clayton of Vanderbilt University, Jeff Wiser of Northrop Grumman, and the anonymous referees for helpful discussions and recommendations regarding the writing of this manuscript.

REFERENCES

1. Kaiser J. Biobanks: population databases boom, from Iceland to the US. *Science*. 2002;298:1158–1161.
2. Kaye J, Heeney C, Hawkins N, et al. Data sharing and genomics—reshaping scientific practices. *Nat Rev Genet*. 2009;10: 331–335.
3. Piwowar H, Becich M, Bilofsky H, et al. Towards a data sharing culture: recommendations for leadership from academic health centers. *PLoS Med*. 2008;5:e183.
4. Karp D, Carlin S, Cook-Deegan R, et al. Ethical and practical issues associated with aggregating databases. *PLoS Med*. 2008;5:e190.
5. McGuire A, Caulfield T, Cho M. Research ethics and the challenge of whole-genome sequencing. *Nat Rev Genet*. 2008;9(2):152–156.
6. 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*. 2005;12:28–34.
7. 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.
8. Jacobs KB, Yeager M, Wacholder S, et al. A new statistic and its power to infer membership in a genome-wide association study using genotype frequencies. *Nat Genet*. 2009;41:1253–1257.
9. Mailman M, Feolo M, Jin Y, et al. The NCBI database of genotypes and phenotypes. *Nat Genet*. 2007;39:1181–1186.
10. Zerhouni E, Nabel E. Protecting aggregate genomic data. *Science*. 2008;322:44.
11. National Institutes of Health. Final NIH statement on sharing research data. NOT-OD-03-032. Feb 26, 2003.
12. National Institutes of Health. Policy for sharing of data obtained in NIH supported or conducted genome-wide association studies (GWAS). Notice NOT-OD-07-088. August 28, 2007.
13. US Department of Health and Human Services. Standards for privacy of individually identifiable health information, final rule. 45 CFR, Parts 160–164. *Fed Regist*. 2002;67(157):53182–53273.
14. Gostin L, Nass S. Reforming the HIPAA privacy rule: safeguarding privacy and promoting research. *JAMA*. 2009;301:1373–1375.
15. National Association of Health Data Organizations. *NAHDO Inventory of State-wide Hospital Discharge Data Activities*. Falls Church, VA: National Association of Health Data Organizations; 2008.
16. Schoenman J, Sutton J, Kintala S, et al. The value of hospital discharge databases: final report to the Agency for Healthcare Research and Quality under contract number 282-98-0024 (task order number 5). White paper, NORC at the University of Chicago, in cooperation with the National Association of Health data Organizations. 2005. Available at http://www.hcup-us.ahrq.gov/reports/final_report.pdf. Accessed August 8, 2009.
17. Brawley S. Submission and retrieval of an aligned set of nucleic acid sequences. *J Phycol*. 1999;35:433–437.
18. Lin Z, Hewett M, Altman R. Using binning to maintain confidentiality of medical data. *Proc AMIA Symp*. 2002:454–458.
19. Lin Z, Owen A, Altman R. Genetics: genomic research and human subject privacy. *Science*. 2004;305:183.
20. Kargupta H, Datta S, Wang Q, et al. Random-data perturbation

techniques and privacy-preserving data mining. *Knowl Inf Syst*. 2005;7:387–414.

21. Sweeney L. Weaving technology and policy together to maintain confidentiality. *J Law Med Ethics*. 1997;25:98–110.
22. Golle P. Revisiting the uniqueness of simple demographics in the U.S. population. In: *Proceedings of the 2006 ACM Workshop on Privacy in the Electronic Society.* ACM Press, 2006:77–80.
23. Sweeney L. *Uniqueness of Simple Demographics in the US Population. White Paper LIDAP-WP4, Laboratory for International Data Privacy, Carnegie Mellon University*, Pittsburgh, PA, 2000.
24. Loukides G, Denny J, Malin B. Do clinical profiles constitute privacy risks? *Proc AMIA Symp*. 2009: to appear.
25. Malin B. Re-identification of familial database records. *Proc AMIA Symp*. 2006:524–528.
26. 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.
27. Altman R, Klein T. Challenges for biomedical informatics and pharmacogenomics. *Annu Rev Pharmacol Toxicol*. 2002;42:113–133.
28. De Moor G, Claerhout B, De Meyer F. Privacy enhancing techniques: the key to secure communication and management of clinical and genomic data. *Methods Inf Med*. 2003;42:148–153.
29. Dugas M, Schoch C, Schnittger S, et al. Impact of integrating clinical and genetic information. *In Silico Biol*. 2002;2:383–391.
30. Sweeney L. Guaranteeing anonymity when sharing medical data, the Datafly system. *Proc AMIA Symp*. 1997:51–55.
31. Malin B. Betrayed by my shadow: learning data identity via trail matching. *Journal of Privacy Technology*. 2005;20050609001.
32. Clabby C. DNA research commons scaled back. *American Scientist*. 2009;97(2):113.
33. Ferris N. The search for John Doe. *Government Health IT Magazine*. 2009.
34. Patoine B. Nervecenter: speed bump for open access to genomic data. *Ann Neurol*. 2008;64:A16–A17.
35. Aldhous P. Genetic data withdrawn amid privacy concerns. *New Sci*. 2008.
36. Sankararaman S, Obozinksi G, Jordan M, et al. Genomic privacy and limits of individual detection in a pool. *Nat Genet*. 2009;41:965–967.
37. GAIN Collaborative Research Group, Manolio T, Rodriguez LL, et al. New models of collaboration in genome-wide association studies: the Genetic Association Information Network. *Nat Genet*. 2007;39: 1045–1051.
38. Manolio T. Collaborative genome-wide association studies of diverse diseases: programs of the NHGRI's office of population genomics. *Pharmacogenomics*. 2009;10:235–241.
39. Greely H. The uneasy ethical and legal underpinnings of large-scale genomic biobanks. *Annu Rev Genomics Hum Genet*. 2007;8:343–364.
40. Caulfield T, Upshur R, Daar A. DNA databanks and consent: a suggested policy option involving an authorization model. *BMC Med Ethics*. 2003;4:1.
41. Sweeney L. K-anonymity: a model for protecting privacy. *Int J Uncertainty Fuzziness Knowl-based Syst*. 2002;10:557–570.