CONCEPTS AND METHODS FOR DE-IDENTIFYING CLINICAL TRIAL DATA

Khaled El Emam, Ph.D. (University of Ottawa) and

Bradley Malin, Ph.D. (Vanderbilt University)

Disclaimer: The authors are responsible for the content of this article, which does not necessarily represent the views of the Institute of Medicine.

Disclosures: The authors report no significant conflicts or financial disclosures related to this work.

Funding: This work was commissioned by the Institute of Medicine Committee on Strategies for Responsible Sharing of Clinical Trial Data.

INTRODUCTION

Context

Very detailed health information about participants is collected during clinical trials. A number of different stakeholders would typically have access to individual-level participant data (IPD), including the study sites, the sponsor of the study, statisticians, Institutional Review Boards (IRBs), and regulators. By IPD we mean individual-level data on trial participants, which is more than the information that is typically included, for example, in clinical study reports (CSRs).

There is increasing pressure to share IPD more broadly than occurs at present. There are many reasons for such sharing, such as transparency in the trial and wider disclosure of adverse events that may have transpired, or to facilitate the reuse of such data for secondary purposes, specifically in the context of health research (Gøtzsche, 2011; IOM, 2013; Vallance and Chalmers, 2013). Many funding agencies tasked with the oversight of research, as well as its funding, are requiring that data collected by the projects they support be made available to others (MRC, 2011; NIH, 2003; Wellcome Trust, 2011). There are current efforts by regulators, such as the European Medicines Agency (EMA, 2014a,b), to examine how to make IPD from clinical trials shared more widely (IOM, 2013). In many cases, however, privacy concerns have been stated as a key obstacle to making these data available (Castellani, 2013; IOM, 2013).

One way in which privacy issues can be addressed is through the protection of the identities of the corresponding research participants. Such “de-identified” or “anonymized” health data (the former term being popular in North America, and the latter in Europe and other

regions) are often considered to be sufficiently devoid of personal health information in many jurisdictions around the world. As such, many privacy laws allow the data to be used and disclosed for any secondary purposes with participant consent. As long as the data are appropriately de-identified, many privacy concerns associated with data sharing can be readily addressed.

It should be recognized that de-identification is not, by any means, the only privacy concern that needs to be addressed when sharing clinical trial data. In fact, there must be a level of governance in place to ensure that the data will not be analyzed or used to discriminate against or stigmatize the participants or certain groups (e.g., religious or ethnic) associated with the study. This is because discrimination and stigmatization can occur even if the data are de-identified.

This paper describes a high-level risk-based methodology that can be followed to de-identify clinical trial IPD. To contextualize our review and analysis of de-identification, we also touch upon additional governance mechanisms, but we acknowledge that a complete treatment of governance is beyond the scope of this paper. Rather, the primary focus here is only on the privacy protective elements.

Data Recipients, Sponsors, and Adversaries

Clinical trial data may be disclosed by making them completely public or through a request mechanism. The data recipient may be a qualified investigator (QI) who must meet specific criteria. There may be other data recipients who are not QIs as well. If the data are made publicly available with no restrictions, however, then other types of users may access the data,

such as journalists and nongovernmental organizations (NGOs). In our discussions we refer to the data recipient as the QI as a primary exemplar, although this is not intended to exclude other possible data recipients (it does make the presentation less verbose).

Data are being disclosed to the QI by the sponsor. We use the term "sponsor" generally to refer to all data custodians who are disclosing IPD, recognizing that the term may mean different entities depending on the context. It may not always be the case that the sponsor is a pharmaceutical company or a medical device company. For example, a regulator may decide to disclose the data to a QI, or a pharmaceutical company may provide the data to an academic institution, whereby that institution becomes the entity that discloses the data.

The term "adversary" is often used in the disclosure control literature to refer to the role of the individual or entity that is trying to re-identify data subjects. Other terms used are "attacker" and "intruder." Discussions about the QI being a potential adversary are not intended to paint QIs as having malicious objectives. Rather, in the context of a risk assessment, one must consider a number of possible data recipients as being potential adversaries and manage the re-identification risk accordingly.

Data Sharing Models

A number of different ways to provide access to IPD have been proposed and used, each with different advantages and risks (Mello et al., 2013). First, there is the traditional public data release where anyone can get access to the data with no registration or conditions. Examples of such releases include the publicly available clinical trial data from the International Stroke Trial

(IST) (Sandercock et al., 2011) and data posted to the Dryad online open access data repository (Dryad, undated; Haggie, 2013).

A second form of data sharing, which is more restrictive, occurs when there exists a formal request and approval process to obtain access to clinical trial data, such as the GlaxoSmithKline (GSK) trials repository (Harrison, 2012; Nisen and Rockhold, 2013); Project Data Sphere (whose focus is on oncology trial data) (Bhattacharjee, 2012; Hede, 2013); the Yale Open Data Access (YODA) Project, which is initially making trial data from Medtronic available (CORE, 2014; Krumholz and Ross, 2011); and the Immunology Database and Analysis Portal (Immport), which is restricted to researchers funded by the Division of Allergy, Immunology, and Transplantation of the National Institute of Allergy and Infectious Diseases (DAIT/NIAID), other approved life science researchers, National Institutes of Health employees, and other preauthorized government employees (ImmPort, undated). More recently, pharmaceutical companies have created the clinicalstudydatarequest.com website, which facilitates data requests to multiple companies under one portal. Following this restrictive model, a request can be processed by the study sponsor or by a delegate of the sponsor (e.g., an academic institution).

A hybrid of the above approaches is a quasi-public release where the data user must agree to some terms of use or sign a "click-through" contract. Click-through contracts are online terms of use that may place restrictions on what can be done with the data and how the data are handled. Regardless, anyone can still download such data. For example, public analytics competition data sets, such as the Heritage Health Prize (El Emam et al., 2012), and data-centric software application development competitions, such as the Cajun Code Fest (Center for

Business and Information Technologies, 2013), fall into this category. In practice, however, click-through terms are not common for the sharing of clinical trial IPD.[1]

A form of data access that does not require any data sharing is when analysts request that the data controller perform an analysis on their behalf. Since this does not involve the sharing of IPD, it is a scenario that we do not consider further in this paper.

Data Sharing Mechanisms

Different mechanisms can be used to share IPD. Clinical trial IPD can be shared either as *microdata* or through an *online portal*. The term "microdata" is commonly used in the disclosure control literature to refer to individual-level raw data (Willenborg and de Waal, 1996, 2001). These microdata may be in the form of one or more flat files or relational databases.

When disclosed as microdata, the data are downloaded as a raw data file that can be analyzed by QIs on their own machines, using their own software if they so wish to do so. The microdata can be downloaded through a website, sent to the QI on a disc, or transferred electronically. If access is through a website, the QI may have to register, sign a contract, or go through other steps before downloading the data.

When a portal is used, the QI can access the data only through a remote computer interface, such that the raw data reside on the sponsor's computers, and all analysis performed is on the sponsor's computers. Data users do not download any microdata to their own local computers through this portal. Under this model, all actions can be audited.

[1] Although the EMA has recently proposed using an online portal to share CSRs using a simple terms-of-use setup, this was not intended to apply to IPD.

A public online portal allows anyone to register and get access to the IPD. Otherwise, the access mechanism requires a formal request process.

De-identification is relevant in both of the aforementioned scenarios. When data are provided as microdata, the de-identification process ensures that each record is protected from the QI and his/her staff as the potential adversary. When data are shared through the portal, a QI or his/her staff may inadvertently recognize a data subject because that data subject is a neighbor, relative, coworker, or famous person (see Box 1).

BOX 1

Types of Re-identification Attacks

For public data, the sponsor needs to make a worst-case assumption and protect against an adversary who is targeting the data subjects with the highest risk of re-identification.

For a nonpublic data set, we consider three types of attacks:

- a deliberate re-identification by the data recipient (or his/her staff and subcontractors);
- an inadvertent re-identification by the data recipient (or his/her staff and subcontractors); and
- a data breach, where data are accidentally exposed to a broader audience.

These three cases are relevant when microdata are being disclosed. If the data are made available through a portal, we assume that the sponsor will ensure that stringent controls and appropriate auditing are in place, which manages risks from the first and third types of attack. In such a case, the second type of attack, where data may be inadvertently re-identified, becomes the primary risk that needs to be managed. An example is if the statistician working with the data inadvertently recognizes someone he or she knows.

The different approaches for sharing clinical trial IPD are summarized in Figure 1.

	Microdata	Online Portal
Public	LEAST CONTROL BY SPONSOR LIMIT CONSTRAINTS ON QI	
Formal Request		MOST CONTROL BY SPONSOR SIGNIFICANT CONSTRAINTS ON QI
	Risks • Deliberate re-identification • Inadvertent re-identification • Accidental release and re-identification	**Risks** • Inadvertent re-identification

FIGURE 1 Different approaches for sharing clinical trial data.

Scope of Data to Be De-identified

It is important to make a distinction between biological, and particularly genomic, data and other types of data. Many clinical trials are creating biorepositories. These may have a pseudonym or other unique identifier for the participant, and a sample or data. The de-identification methods we describe in this paper are applicable to clinical, administrative, and survey data. Genomic data raise a different set of issues. These issues are addressed directly in a later section of this paper.

Clinical trial data can be shared at multiple levels of detail. For example, the data can be raw source data or analysis-ready data. We assume that the data are analysis-ready and that no data cleansing is required before de-identification.

Existing Standards for De-identification

Various regulations associated with data protection around the world permit the sharing of de-identified (or similarly termed) data. For instance, EU Data Protection Directive 95/46/EC, which strictly prohibits secondary uses of person-specific data without individual consent, provides an exception to the ruling in Recital 26, which states that the "principles of protection shall not apply to data rendered anonymous in such a way that the data subject is no longer identifiable." However, what does it mean for data to be "identifiable"? How do we know when they are no longer identifiable? The Data Protection Directive, and similar directives around the world, do not provide explicit guidelines regarding how data should be protected. An exception to this rule is a code of practice document published by the U.K. Information Commissioner's Office (ICO) (ICO, 2012). And while this document provides examples of de-identification methods and issues to consider when assessing the level of identifiability of data, it does not provide a full methodology or specific standards to follow.

There are, however, de-identification standards provided in the Privacy Rule of the U.S. Health Insurance Portability and Accountability Act of 1996 (HIPAA) and subsequent guidance published by the Office for Civil Rights (OCR) at the U.S. Department of Health and Human Services (HHS) (HHS, 2012). This rule is referred to by many regulatory frameworks around the world, and the principles are strongly related to those set forth in the United Kingdom's code of practice document mentioned above.

Two of the key existing standards for the de-identification of health microdata are described in the HIPAA Privacy Rule. It should be recognized that HIPAA applies only to "covered entities" (i.e., health plans, health care clearinghouses, and health care providers that transmit health information electronically) in the United States. It is likely that in many instances, the sponsors of clinical trials will not fall into this class. However, these de-

identification standards have been in place for approximately a decade, and there is therefore a considerable amount of real-world experience in their application. They can serve as a good launching point for examining best practices in this area. For the disclosure of clinical trial data, the HIPAA Privacy Rule de-identification standards offer a practically defensible foundation even if they are not a regulatory requirement.

According to section 164.514 of the HIPAA Privacy Rule, "health information that does not identify an individual and with respect to which there is no reasonable basis to believe that the information can be used to identify an individual is not individually identifiable health information." Section 164.514(b) of the Privacy Rule contains the implementation specifications that a covered entity, or affiliated business associate, must follow to meet the de-identification standard. In particular, the Privacy Rule outlines two routes by which health data can be designated as de-identified. These are illustrated in Figure 2.

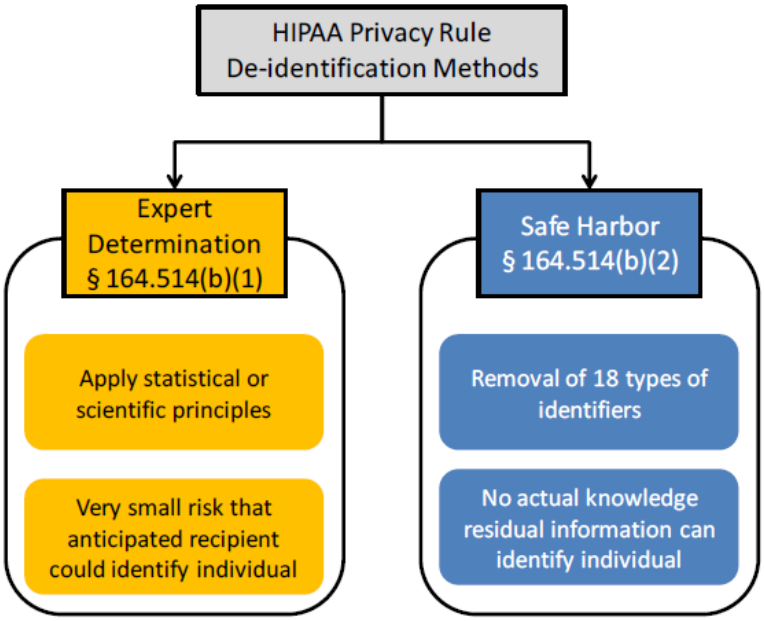

FIGURE 2 The two de-identification standards in the HIPAA Privacy Rule.
SOURCE: Reprinted from a document produced by OCR (HHS, 2012).

The first route is the "Safe Harbor" method. Safe Harbor requires the manipulation of 18 fields in the data set as described in Box 2. The Privacy Rule requires that a number of these data elements be "removed." However, there may be acceptable alternatives to actual removal of values as long as the risk of reverse engineering the original values is very small. Compliance with the Safe Harbor standard also requires that the sponsor not have any actual knowledge that a data subject can be re-identified. Assumptions of the Safe Harbor method are listed in Box 3.

BOX 2

The Safe Harbor De-identification Standard

1. Names;
2. All geographic subdivisions smaller than a state, including street address, city, county, precinct, zip code, and their equivalent geocodes, except for the initial three digits of a zip code if, according to the current publicly available data from the Bureau of the Census:
    a) The geographic unit formed by combining all zip codes with the same three initial digits contains more than 20,000 people; and
    b) The initial three digits of a zip code for all such geographic units containing 20,000 or fewer people is changed to 000.
3. All elements of dates (except year) for dates directly related to an individual, including birth date, admission date, discharge date, date of death; and all ages over 89 and all

elements of dates (including year) indicative of such age, except that such ages and elements may be aggregated into a single category of age 90 or older;

4. Telephone numbers;
5. Fax numbers;
6. Electronic mail addresses;
7. Social security numbers;
8. Medical record numbers;
9. Health plan beneficiary numbers;
10. Account numbers;
11. Certificate/license numbers;
12. Vehicle identifiers and serial numbers, including license plate numbers;
13. Device identifiers and serial numbers;
14. Web universal resource locators (URLs);
15. Internet protocol (IP) address numbers;
16. Biometric identifiers, including finger and voice prints;
17. Full face photographic images and any comparable images; and
18. Any other unique identifying number, characteristic, or code.

BOX 3

Assumptions of the HIPAA Safe Harbor Method

- There are only two quasi-identifiers that need to be manipulated in a data set: dates and zip codes.
- The adversary does not know who is in the data set (i.e., would not know which individuals participated in the clinical trial).
- All dates are quasi-identifiers.

While the application of Safe Harbor is straightforward, however, there are clearly instances in which dates and more fine-grained geographic information are necessary. In practice the Safe Harbor standard would remove critical geospatial and temporal information from the data (see items 2 and 3 in Box 2), potentially reducing the utility of the data. Many meaningful analyses of clinical trial data sets require the dates and event order to be clear. For example, in a Safe Harbor data set, it would not be possible to include the dates when adverse events occurred.

In recognition of the limitations of de-identification via Safe Harbor, the HIPAA Privacy Rule provides for an alternative in the form of the Expert Determination method. This method has three general requirements:

- The de-identification must be based on *generally accepted statistical and scientific principles and methods for rendering information not individually identifiable*. This means that the sponsor needs to ensure that there is a body of work that justifies and evaluates the methods that are used for the de-identification, and that these methods must be generally known (i.e., undocumented methods or proprietary methods that have never been published would be difficult to classify as "generally accepted").

- The risk of re-identification needs to be *very small* such that the information could not be used, alone or in combination with other reasonably available information, by an anticipated recipient to identify an individual who is a subject of the information. However, the mechanism for measuring re-identification risk is not defined in the HIPAA Privacy Rule, and what would be considered *very small* risk also is not defined. Therefore, the de-identification methodology must include some manner of measuring re-identification risk in a defensible way, and have a repeatable process to follow that allows for the definition of *very small* risk.
- Finally, the *methods and results of the analysis that justify such determination* must be documented. While the basic principles of de-identification are expected to be consistent across all clinical trials, the details will be different for each study, and these details also need to be documented.

These conditions are reasonable for a de-identification methodology and are consistent with the guidance that has been produced by other agencies and regulators (Canadian Institute for Health Information, 2010; ICO, 2012). They also serve as a set of conditions that must be met for the methods described here.

Unique and Derived Codes under HIPAA

According to the 18th item in Safe Harbor (see Box 2), "any unique identifying number, characteristic, or code" must be removed from the data set; otherwise it would be considered personal health information. However, in lieu of removing the value, it may be hashed or encrypted. This would be called a "pseudonym." For example, the unique identifier may be a

participant's clinical trial number, and this is encrypted with a secret key to create a pseudonym. A similar scheme for creating pseudonyms would be used under the Expert Determination method.

However, in the HIPAA Privacy Rule at § 164.514(c), it is stated that any code that is *derived* from information about an individual is considered identifiable data. However, such pseudonyms are practically important for knowing which records belong to the same clinical trial participant and constructing the longitudinal record of a data subject. Not being able to create derived pseudonyms means that random pseudonyms must be created. To be able to use random pseudonyms, one must maintain a crosswalk between the individual identity and the random pseudonym. The crosswalk allows the sponsor to use the same pseudonym for each participant across data sets and to allow re-identification at a future date if the need arises. These crosswalks, which are effectively linking tables between the pseudonym and the information about the individual, arguably present an elevated privacy risk because clearly identifiable information must now be stored somehow. Furthermore, the original regulations did not impose any controls on this crosswalk table.

For research purposes, the Common Rule will also apply. Under the Common Rule, which guides IRBs, if the data recipient has no means of getting the key, for example, through an agreement with the sponsor prohibiting the sharing of keys under any circumstances or through organizational policies prohibiting such an exchange, then creating such derived pseudonyms is an acceptable approach (HHS, 2004, 2008b).

Therefore, there is an inconsistency between the Privacy Rule and the Common Rule in that the former does not permit derived pseudonyms, while the latter does. This is well documented (Rothstein, 2005, 2010). However, in the recent guidelines from OCR, this is

clarified to state that "a covered entity may disclose codes derived from PHI as part of a de-identified data set if an expert determines that the data meets the de-identification requirements at §164.514(b)(1)" (HHS, 2012). This means that a derived code, such as an encryption or hash function, can be used as a pseudonym as long as there is assurance that the means to reverse that pseudonym are tightly controlled. There is now clarity and consistency among rules in that if there is a defensible mechanism whereby reverse engineering a derived pseudonym has a very small probability of being successful, this is permitted.

Is it Necessary to Destroy Original Data?

Under the Expert Determination method, the re-identification risk needs to be managed assuming that the adversary is "an anticipated recipient" of the data. This limits the range of adversaries that needs to be considered because in our context, the anticipated recipient is the QI.

However, under the EU Data Protection Directive, the adversary may be the "data controller or any other person." The data controller is the sponsor or the QI receiving the de-identified data. There are a number of challenges with interpreting this at face value.

One practical issue is that the sponsor will, by definition, be able to re-identify the data because the sponsor will retain the original clinical trial data set. The Article 29 Working Party has proposed that, effectively, the sponsor needs to destroy or aggregate the original data to be able to claim that the data provided to the QI are truly de-identified (Article 29 Data Protection Working Party, 2014). This means that the data are not de-identified if there exists another data set that can re-identify it, even in the possession of another data controller. Therefore, because the identified data exist with the sponsor, the data provided to the QI cannot be considered de-identified. This is certainly not practical because the original data are required for legal reasons

(e.g., clinical trial data need to be retained for an extended period of time whose duration depends on the jurisdiction). Such a requirement would discourage de-identification by sponsors and push them to share identifiable data, which arguably would increase the risk of re-identification for trial participants significantly.

In an earlier opinion the Article 29 Data Protection Working Party (2007) emphasized the importance of "likely reasonable" in the definition of identifiable information in the 95/46/EC Directive. In that case, if it is not likely reasonable that data recipients would be able to readily re-identify the anonymized data because they do not have access to the original data, those anonymized data would not be considered personal information. That would seem to be a more reasonable approach that is consistent with interpretations in other jurisdictions.

Is De-identification a Permitted Use?

Retroactively obtaining participant consent to de-identify data and use them for secondary analysis may introduce bias in the data set (El Emam, 2013). If de-identification is a permitted use under the relevant regulations, then de-identification can proceed without seeking participant consent. Whether that is the case will depend on the prevailing jurisdiction.

Under HIPAA and extensions under the Health Information Technology for Economic and Clinical Health (HITECH) Omnibus Rule, de-identification is a permitted use by a covered entity. However, a business associate can de-identify a data set only if the business associate agreement explicitly allows for that. Silence on de-identification in a business associate agreement is interpreted as not permitting de-identification.

In other jurisdictions, such as Ontario, the legislation makes explicit that de-identification is a permitted use (Perun et al., 2005).

Terminology

Terminology in this area is not always clear, and different authors and institutions use the same terms to mean different things or different terms to mean the same thing (Knoppers and Saginur, 2005). Here, we provide the terminology and definitions used in this paper.

The International Organization for Standardization (ISO) Technical Specification on the pseudonymization of health data defines relevant terminology for our purposes. The term "anonymization" is defined as a "process that removes the association between the identifying data set and the data subject" (ISO, 2008). This is consistent with current definitions of "identity disclosure," which corresponds to assigning an identity to a data subject in a data set (OMB, 1994; Skinner, 1992). For example, an identity disclosure would transpire if the QI determined that the third record (ID = 3) in the example data set in Table 1 belonged to Alice Brown. Thus, anonymization is the process of reducing the probability of identity disclosure to a very small value.

TABLE 1 An Example of Data Used to Illustrate a Number of Concepts Referred to Throughout This Paper

Quasi-identifiers			Other Variables	
ID	Sex	Year of Birth	Lab Test	Lab Result
1	Male	1959	Albumin, Serum	4.8
2	Male	1969	Creatine kinase	86
3	Female	1955	Alkaline Phosphatase	66
4	Male	1959	Bilirubin	Negative
5	Female	1942	BUN/Creatinine Ratio	17
6	Female	1975	Calcium, Serum	9.2

7	Female	1966	Free Thyroxine Index	2.7
8	Female	1987	Globulin, Total	3.5
9	Male	1959	B-type natriuretic peptide	134.1
10	Male	1967	Creatine kinase	80
11	Male	1968	Alanine aminotransferase	24
12	Female	1955	Cancer antigen 125	86
13	Male	1967	Creatine kinase	327
14	Male	1967	Creatine kinase	82
15	Female	1966	Creatinine	0.78
16	Female	1955	Triglycerides	147
17	Male	1967	Creatine kinase	73
18	Female	1956	Monocytes	12
19	Female	1956	HDL Cholesterol	68
20	Male	1978	Neutrophils	83
21	Female	1966	Prothrombin Time	16.9
22	Male	1967	Creatine kinase	68
23	Male	1971	White Blood Cell Count	13.0
24	Female	1954	Hemoglobin	14.8
25	Female	1977	Lipase, Serum	37
26	Male	1944	Cholesterol, Total	147
27	Male	1965	Hematocrit	45.3

Arguably, the term “anonymization” would be the appropriate term to use here given its more global utilization. However, to remain consistent with the HIPAA Privacy Rule, we use the term “de-identification” in this paper.

Beyond identity disclosure, organizations (and privacy professionals) are, at times, concerned about “attribute disclosure” (OMB, 1994; Skinner, 1992). This occurs when a QI learns a sensitive attribute about a participant in the database with a sufficiently high probability, even if the Q1 does not know which specific record belongs to that patient (Machanavajjhala et al., 2007; Skinner, 1992). For example, in Table 1, all males born in 1967 had a creatinekinease lab test. Assume that an adversary does not know which record belongs to Almond Zipf (who has record ID = 17; see Table 2). However, since Almond is male and was born in 1967, the QI will discover something new about him—that he had a test often administered to individuals showing symptoms of a heart attack. All known re-identification attacks are identity disclosures

and not attribute disclosures (El Emam et al., 2011a).[2] Furthermore, privacy statutes and regulations in multiple jurisdictions, including the HIPAA Privacy Rule, the Ontario Personal Health Information Act (PHIPA), and the EU Data Protection Directive, consider identity disclosure only in their definitions of personal health information. While participants may consider certain types of attribute disclosure to be a privacy violation, it is not considered so when the objective is anonymization of the data set.

TABLE 2 Identities of Participants from the Hypothetical Data Set

ID	Name
1	John Smith
2	Alan Smith
3	Alice Brown
4	Hercules Green
5	Alicia Freds
6	Gill Stringer
7	Marie Kirkpatrick
8	Leslie Hall
9	Douglas Henry
10	Fred Thompson
11	Joe Doe
12	Lillian Barley
13	Deitmar Plank
14	Anderson Hoyt
15	Alexandra Knight
16	Helene Arnold
17	Almond Zipf
18	Britney Goldman
19	Lisa Marie
20	William Cooper
21	Kathy Last
22	Deitmar Plank
23	Anderson Hoyt
24	Alexandra Knight
25	Helene Arnold
26	Anderson Heft
27	Almond Zipf

[2] This statement does not apply to genomic data. See the summary of evidence on genomic data later in this paper for more detail.

Technical methods have been developed to modify the data to protect against attribute disclosure (Fung et al., 2010). However, these methods have rarely, if ever, been used in practice for the disclosure of health data. One possible reason for this is that they distort the data to such an extent that the data are no longer useful for analysis purposes. There are other, nontechnical approaches that are more appropriate for addressing the risks of attribute disclosure, and in the final section on governance we provide a description of how a sponsor can protect against attribute disclosure. Therefore, our focus in this paper is on identity disclosure.

HOW TO MEASURE THE RISK OF RE-IDENTIFICATION

We begin with some basic definitions that are critical for having a meaningful discussion about how re-identification works. Along the way, we address some of the controversies around de-identification that have appeared in the literature and the media.

Categories of Variables

It is useful to differentiate among the different types of variables in a clinical trial data set. The way the variables are handled during the de-identification process will depend on how they are categorized. We make a distinction among three types of variables (Samarati, 2001; Sweeney, 2002):

- **Directly identifying variables.** Direct identifiers have two important characteristics: (1) one or more direct identifiers can be used to uniquely identify an individual, either

by themselves or in combination with other readily available information; and (2) they often are not useful for data analysis purposes. Examples of directly identifying variables include names, email address, and telephone numbers of participants. It is uncommon to perform data analysis on clinical trial participant names and telephone numbers.

- **Indirectly identifying variables (quasi-identifiers).** Quasi-identifiers are the variables about research participants in the data set that a QI can use, either individually or in combination, to re-identify a record. If an adversary does not have background knowledge of a variable, it cannot be a quasi-identifier. The means by which an adversary can obtain such background knowledge will determine which attacks on a data set are plausible. For example, the background knowledge may be available because the adversary knows a particular target individual in the disclosed clinical trial data set, an individual in the data set has a visible characteristic that is also described in the data set, or the background knowledge exists in a public or semipublic registry. Examples of quasi-identifiers include sex, date of birth or age, locations (such as postal codes, census geography, and information about proximity to known or unique landmarks), language spoken at home, ethnic origin, aboriginal identity, total years of schooling, marital status, criminal history, total income, visible minority status, activity difficulties/reductions, profession, event dates (such as admission, discharge, procedure, death, specimen collection, visit/encounter), codes (such as diagnosis codes, procedure codes, and adverse event codes), country of birth, birth weight, and birth plurality.

- **Other variables.** These are the variables that are not really useful for determining an individual's identity. They may or may not be clinically relevant.

Individuals can be re-identified because of the directly identifying variables and the quasi-identifiers. Therefore, our focus is on these two types of variables.

Classifying Variables

An initial step in being able to reason about the identifiability of a clinical trial data set is to classify the variables into the above categories. We consider the process for doing so below.

Is It an Identifier?

There are three conditions for a field to be considered an identifier (of either type). These conditions were informed by HHS's de-identification guidelines (HHS, 2012).

Replicability

The field values must be sufficiently stable over time so that the values will occur consistently in relation to the data subject. For example, the results of a patient's blood glucose level tests are unlikely to be replicable over time because they will vary quite a bit. If a field value is not replicable, it will be challenging for an adversary to use that information to re-identify an individual.

Distinguishability

The variable must have sufficient variability to distinguish among individuals in a data set. For example, in a data set of only breast cancer patients, the diagnosis code (at least at a high level) will have little variation. On the other hand, if a variable has considerable variation among the data subjects, it can distinguish among individuals more precisely. That diagnosis field will be quite distinguishable in a general insurance claims database.

Knowability

An adversary must know the identifiers about the data subject in order to re-identify them. If a variable is not knowable by an adversary, it cannot be used to launch a re-identification attack on the data.

When we say that a variable is knowable, it also means that the adversary has an identity attached to that information. For example, if an adversary has a zip code and a date of birth, as well as an identity associated with that information (such as a name), then both the zip code and date of birth are knowable.

Knowability will depend on whether an adversary is an acquaintance of a data subject. If the adversary is an acquaintance, such as a neighbor, coworker, relative, or friend, it can be assumed that certain things will be known. Things known by an acquaintance will be, for example, the subject's demographics (e.g., date of birth, gender, ethnicity, race, language spoken at home, place of birth, and visible physical characteristics). An acquaintance may also know some socioeconomic information, such as approximate years of education, approximate income, number of children, and type of dwelling.

A nonacquaintance will know things about a data subject in a number of different ways, in decreasing order of likelihood:

- The information can be inferred from other knowable information or other variables that determined to be identifiers. For example, birth weight can often be inferred from weeks of gestation. If weeks of gestation are included in the database, birth weight can be determined with reasonable accuracy.
- The information is publicly available. For example, the information is in a public registry, or it appears in a newspaper article (say, an article about an accident or a famous person). Information can also become public if self-revealed by individuals. Examples are information posted on social networking sites and broadcast email announcements (e.g., births). It should be noted that only information that many people would self-reveal should be considered an identifier. If there is a single example or a small number of examples of people who are revealing everything about their lives (e.g., a quantified-self enthusiast who is also an exhibitionist), this does not mean that this kind of information is an identifier for the majority of the population.
- The information is in a semipublic registry. Access to these registries may require a nominal fee or application process.
- The information can be purchased from commercial data brokers. Use of commercial databases is not inexpensive, so an adversary would need to have a strong motive to use such background information.

Some of these data sources can be assessed objectively (e.g., whether there is relevant public information). In other cases, the decision will be subjective and may vary over time.

A Suggested Process for Determining Whether a Variable Is an Identifier

A simple way to determine whether a variable is an identifier is to ask an expert, internal or external to the sponsor, to do so. There are other, more formal processes that can be used as well.

There are two general approaches to classifying variables. In one approach, two analysts who know the data and the data subject population classify the variables independently; then some measure of agreement is computed. A commonly used measure of agreement is Cohen's Kappa (Cohen, 1960). If this value is above 0.8, there is arguably general consensus, and the two analysts will meet to resolve the classifications on which they had disagreements. The results of this exercise are then retained as documentation.

If the Kappa value is less than 0.8, there is arguably little consensus. In such a case, it is recommended that a group of individuals at the sponsor site review the field classifications and reach a classification consensus. This consensus then needs to be documented, along with the process used to reach it. This process provides the data custodian with a defensible classification of variables.

Is It a Direct or Indirect Identifier?

Once a variable has been determined to be an identifier, it is necessary to determine whether it is a direct or indirect (quasi-) identifier. If the field uniquely identifies an individual

(e.g., a social security number), it will be treated as a direct identifier. If it is not unique, the next question is whether it is likely to be used for data analysis. If so, it should be treated as a quasi-identifier. This is an important decision because the techniques often used to protect direct identifiers distort the data and their truthfulness significantly.

Is it possible to know which fields will be used for analysis at the time that de-identification is being applied? In many instances, an educated judgment can be made, for example, about potential outcome variables and confounders.

The overall decision rule for classifying variables is shown in Figure 3.

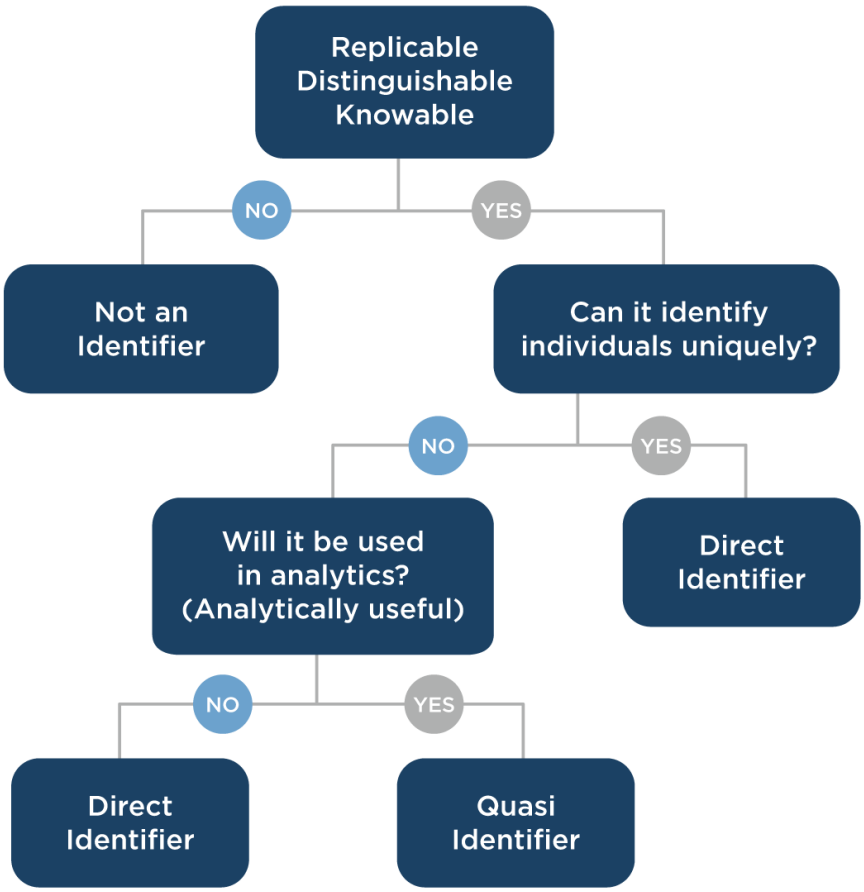

FIGURE 3 Decision rule for classifying identifiers.
SOURCE: Reprinted with permission from El Emam and colleagues (2014).

How Is Re-identification Probability Measured?

Measurement of re-identification risk is a topic that has received extensive study over multiple decades. We examine it at a conceptual level to illustrate key concepts. This discussion builds on the classification of variables described above.

The Risk of Re-identification for Direct Identifiers

We define risk as the probability of re-identifying a trial participant. In practice, we consider the risk of re-identification for direct identifiers to be 1. If a direct identifier does exist in a clinical trial data set, then by definition it will be considered to have a very high risk of re-identification.

Strictly speaking, the probability is not always 1. For example, consider the direct identifier "Last Name." If a trial participant is named "Smith," it is likely that there are other people in the trial named "Smith," and this is even more likely in the community where that participant lives. However, assuming that the probability of re-identification is equal to 1 is a simplification that has little impact in practice, errs on the conservative side, and makes it possible to focus attention on the quasi-identifiers, which is where, in many instances, the most data utility lies.

Two methods can be applied to protect direct identifiers. The first is suppression, or removal of the variable. For example, when a clinical trial data set is disclosed, all of the names of the participants are stripped from the data set. The second method is to create a pseudonym

(ISO, 2008). Pseudonymization is also sometimes called "coding" in the health research literature (Knoppers and Saginur, 2005).[3] There are different schemes and technical methods for pseudonymization, such as single and double coding, reversible or irreversible pseudonyms, and encryption and hashing techniques. If executed well, pseudonymization ensures that the probability of re-identification is very small. There is no need to measure this probability on the data after suppression or pseudonymization because in almost all cases, that value is going to be very small.

Quasi-identifiers, however, cannot be protected using such procedures. This is because the resulting data, in almost all cases, will not be useful for analytic purposes. Therefore, a different set of approaches is required for measuring and de-identifying quasi-identifiers.

The Risk of Re-identification for Quasi-identifiers

Equivalence Classes

All the records that share the same values on a set of quasi-identifiers are called an "equivalence class." For example, consider the quasi-identifiers in Table 1—sex and age. All the records in Table 1 for males born in 1967 (i.e., records 10, 13, 14, 17, and 22) form an equivalence class. Equivalence class sizes for a data concept, such as age, potentially change during de-identification. For example, there may be five records for males born in 1967. When the precision of age is reduced to a 5-year interval, there are eight records for males born between 1965 and 1969 (i.e., records 2, 10, 11, 13, 14, 17, 22, and 27). In general, there is a

[3] A case can made for just using the term "coding" rather than the term "pseudonymization" because it is easier to remember and pronounce. That is certainly a good reason to use the former term as long as the equivalence of the two terms is noted, since "pseudonymization" is the term used in an ISO technical specification.

trade-off between the level of detail provided for a data concept and the size of the corresponding equivalence classes, with more detail being associated with smaller equivalence classes.

The most common way to measure the probability of re-identification for a record in a data set is for the probability to be equal to 1 divided by the size of its equivalence class. For example, record number 14 is in an equivalence class of size five, and therefore its probability of re-identification is 0.2. Record number 27 is in an equivalence class of size one and therefore its probability of re-identification is equal to 1 divided by 1. Records that are in equivalence classes of size one are called "uniques." In Table 3, we have assigned the probability to each record in our example.

TABLE 3 The Data Set in Table 1 with the Probabilities of Re-identification per Record Added

	Quasi-identifiers			Probability of Re-identification
ID	Sex	Year of Birth		
1	Male	1959	…	0.33
2	Male	1969	…	1
3	Female	1955	…	0.33
4	Male	1959	…	0.33
5	Female	1942	…	1
6	Female	1975	…	1
7	Female	1966	…	0.33
8	Female	1987	…	1
9	Male	1959	…	0.33
10	Male	1967	…	0.2
11	Male	1968	…	1
12	Female	1955	…	0.33
13	Male	1967	…	0.2
14	Male	1967	…	0.2
15	Female	1966	…	0.33
16	Female	1955	…	0.33
17	Male	1967	…	0.2
18	Female	1956	…	0.5
19	Female	1956	…	0.5
20	Male	1978	…	1
21	Female	1966	…	0.33
22	Male	1967	…	0.2
23	Male	1971	…	1

24	Female	1954	…	1
25	Female	1977	…	1
26	Male	1944	…	1
27	Male	1965	…	1

This probability applies under two conditions: (1) the adversary knows someone in the real world and is trying to find the record that matches that individual, and (2) the adversary has selected a record in the data set and is trying to find the identity of that person in the real world. Both of these types of attacks on health data have occurred in practice, and therefore both perspectives are important to consider. An example of the former perspective is when an adversary gathers information from a newspaper and attempts to find the data subject in the data set. An example of the latter attack is when the adversary selects a record in the data set and tries to match it with a record in the voter registration list.

A key observation here is that the probability of re-identification is not based solely on the uniques in the data set. For example, record number 18 is not a unique, but it still has quite a high probability of re-identification. Therefore, it is recommended that the risk of re-identification be considered, and managed, for both uniques and nonuniques.

Maximum Risk

One way to measure the probability of re-identification for the entire data set is through the maximum risk, which corresponds to the maximum probability of re-identification across all records. From Table 3, it can be seen that there is a unique record, such that the maximum risk is 1 for this data set.

Average Risk

The average risk corresponds to the average across all records in the data set. In the example of Table 3, this amounts to 0.59. By definition, the average risk for a data set will be no greater than the maximum risk for the same data set.

Which Risk Metric to Use

As the data set is modified, the risk values may change. For example, consider Table 4, in which year of birth has been generalized to decade of birth. The maximum risk is still 1, but the average risk has declined to 0.33. The average risk will be more sensitive than the maximum risk to modifications to the data.

TABLE 4 The Data Set in Table 1 After Year of Birth Has Been Generalized to Decade of Birth, with the Probabilities of Re-identification per Record Added

	Quasi-identifiers			
ID	Sex	Decade of Birth		Probability of Re-identification
1	Male	1950-1959	…	0.33
2	Male	1960-1969	…	0.125
3	Female	1950-1959	…	0.167
4	Male	1950-1959	…	0.33
5	Female	1940-1949	…	1
6	Female	1970-1979	…	0.33
7	Female	1960-1969	…	0.33
8	Female	1980-1989	…	1
9	Male	1950-1959	…	0.33
10	Male	1960-1969	…	0.125
11	Male	1960-1969	…	0.125
12	Female	1950-1959	…	0.167
13	Male	1960-1969	…	0.125
14	Male	1960-1969	…	0.125
15	Female	1960-1969	…	0.33
16	Female	1950-1959	…	0.167
17	Male	1960-1969	…	0.125
18	Female	1950-1959	…	0.167
19	Female	1950-1959	…	0.167
20	Male	1970-1979	…	1
21	Female	1960-1969	…	0.33
22	Male	1960-1969	…	0.125

23	Male	1970-1979	…	0.33
24	Female	1950-1959	…	0.167
25	Female	1970-1979	…	0.33
26	Male	1940-1949	…	1
27	Male	1960-1969	…	0.125

Since the average risk is no greater than the maximum risk, the latter is generally used when a data set is going to be disclosed publicly (El Emam, 2013). This is because a dedicated adversary who is launching a demonstration attack against a publicly available data set will target the record(s) in the disclosed clinical trial data set with the maximum probability of re-identification. Therefore, it is prudent to protect against such an adversary by measuring and managing maximum risk.

The average risk, by comparison, is more suitable for nonpublic data disclosures. For nonpublic data disclosures, some form of data sharing agreement with prohibitions on re-identification can be expected. In this case, it can be assumed that any data subject may be targeted by the adversary.

As a general rule, it is undesirable to have unique records in the data set after de-identification. In the example of Table 1, there are unique records both in the original data set and after year of birth has been changed to decade of birth (see Table 4). For example, record 26 is unique in Table 4. Unique records have a high risk of re-identification. Also, as a general rule, it is undesirable to have records with a probability of re-identification equal to 0.5 in the data set.

With average risk, one can have data sets with an acceptably small average risk but with unique records or records in equivalence classes of size 2. To avoid that situation, one can use the concept of "strict average risk." Here, maximum risk is first evaluated to ensure that it is at or below 0.33. If that condition is met, average risk is computed. This two-step measure ensures that there are no uniques or doubles in the data set.

In the example data set in Table 4, the strict average risk is 1. This is because the maximum risk is 1, so the first condition is not met. However, the data set in Table 5 has a strict average risk of 0.33. Therefore, in practice, maximum risk or strict average risk would be used to measure re-identification risk.

TABLE 5 The Generalized Data Set with No Uniques or Doubles

	Quasi-identifiers			
ID	Sex	Decade of Birth		Probability of Re-identification
1	Male	1950-1959	…	0.33
2	Male	1960-1969	…	0.125
3	Female	1950-1959	…	0.167
4	Male	1950-1959	…	0.33
6	Female	1970-1979	…	0.33
7	Female	1960-1969	…	0.33
9	Male	1950-1959	…	0.33
10	Male	1960-1969	…	0.125
11	Male	1960-1969	…	0.125
12	Female	1950-1959	…	0.167
13	Male	1960-1969	…	0.125
14	Male	1960-1969	…	0.125
15	Female	1960-1969	…	0.33
16	Female	1950-1959	…	0.167
17	Male	1960-1969	…	0.125
18	Female	1950-1959	…	0.167
19	Female	1950-1959	…	0.167
21	Female	1960-1969	…	0.33
22	Male	1960-1969	…	0.125
23	Male	1970-1979	…	0.33
24	Female	1950-1959	…	0.167
25	Female	1970-1979	…	0.33
27	Male	1960-1969	…	0.125

Samples and Populations

The above examples are based on the premise that an adversary knows who is in the data set. Under those conditions, the manner in which the risk metrics have been demonstrated is correct. We call this a “closed” data set. There are situations in which this premise holds true.

For instance, one such case occurs when the data set covers everyone in the population. A second case is when the data collection method itself discloses who is in the data set. Here are several examples in which the data collection method makes a data set closed:

- If everyone attending a clinic is screened into a trial, an adversary who knows someone who attends the clinic will know that that individual is in the trial database.
- A study of illicit drug use among youth requires parental consent, which means that parents will know if their child is in the study database.
- The trial participants self-reveal that they are taking part in a particular trial, for example, on social networks or on online forums.

If it is not possible to know who is in the data set, the trial data set can be considered to be a sample from some population. We call this an "open" data set. Because the data set is a sample, there is some uncertainty about whether a person is in the data set or not. This uncertainty can reduce the probability of re-identification.

When the trial data set is treated as a sample, the maximum and average risk need to be estimated from the sample data. The reason is that in a sample context, the risk calculations depend on the equivalence class size in the population as well. Therefore, the population equivalence class sizes need to be estimated for the same records. Estimates are needed because in most the cases, the sponsor will not have access to the population data.

There is a large body of work on these estimators in the disclosure control literature (e.g., Dankar et al., 2012; Skinner and Shlomo, 2008). A particularly challenging estimation problem is deciding whether a unique record in the sample is also a unique in the population. If a record is

unique in the sample, it may be because the sampling fraction is so small that all records in the sample are uniques. Yet a record may be unique in the sample because it is also unique in the population.

Under these conditions, appropriate estimators need to be used to compute the maximum and average risk correctly. In general, when the data set is treated as a sample, the probability of re-identification will be no greater than the probability associated with situations in which the data set is not treated as a sample (i.e., the adversary knows who is in the data set).

Re-identification Risk of Participants with Rare Diseases

It is generally believed that clinical trials conducted on rare diseases will always have a high risk of re-identification. It is true that the risk of re-identification will, in general, be higher than that for nonrare diseases. However, it is not necessarily too high. If the data set is open with a small sampling fraction and one is using (strict) average risk, the risk of re-identification may be acceptably small. The exact risk value will need to be calculated on the actual data set to make that determination.

Taking Context into Account

Determining whether a data set is disclosed to the public or a more restricted group of recipients illustrates how context is critical. In the case of the recipient, for instance, it informs us which metric is more appropriate. However, this is only one aspect of the context surrounding a data set, and a more complete picture can be applied to make more accurate assessments of re-identification risk.

For a public data release, we assume that the adversary will launch a demonstration attack, and therefore it is necessary to manage maximum risk. There are no other controls that can be put in place. For a nonpublic data, set we consider three types of attacks that cover the universe of attacks: deliberate, inadvertent, and breach (El Emam, 2013; El Emam and Arbuckle, 2013).

A **deliberate attack** transpires when the adversary deliberately attempts to re-identify individuals in the data set. This may be a deliberate decision by the leadership of the data recipient (e.g., the QI decides to re-identify individuals in order to link to another data set) or by a rogue employee associated with the data recipient. The probability that this type of attack will be successful can be computed as follows:

$$Pr(\text{re-id, attempt}) = Pr(\text{re-id} \mid \text{attempt}) \times Pr(\text{attempt}) \quad (1)$$

where the term Pr(attempt) captures the probability that a deliberate attempt to re-identify the data will be made by the data recipient. The actual value for Pr(attempt) will depend on the security and privacy controls that the data recipient has in place and the contractual controls that are being imposed as part of the data sharing agreement. The second term, Pr(re-id | attempt), corresponds to the probability that the attack will be successful in the event that the recipient has chosen to commit the attack. This conditional can be measured from the actual data.

An **inadvertant attack** transpires when a data analyst working with the QI (or the QI himself/herself) inadvertently re-identifies someone in the data set. For instance, this could occur when the recipient is already aware of the identity of someone in the data set, such as a friend;

relative; or, more generally, an acquaintance. The probability of successful re-identification in this situation can be computed as follows:

$$Pr(\text{re-id, acquaintance}) = Pr(\text{re-id} \mid \text{acquaintance}) \times Pr(\text{acquaintance}) \quad (2)$$

There are defensible ways to compute *Pr*(acquaintance) (El Emam, 2013), which evaluates the probability of an analyst knowing someone in the data set. For example, if the trial is of a breast cancer treatment, then *Pr*(acquaintance) is the probability of the analyst knowing someone who has breast cancer. The value for *Pr*(re-id | acquaintance) needs to be computed from the data. Box 4 considers the question of whether it is always necessary to be concerned about the risk of inadvertent re-identification.

BOX 4

Is It Always Necessary to be Concerned About the Risk of Inadvertent Re-identification?

In the context of data release through an online portal, an argument can be made that the sponsor imposes significant security and privacy controls and requires the QI to sign a contract that contains the relevant prohibitions (e.g., a prohibition on re-identification attacks). This means that the probability of re-identification under these two conditions is likely to be very small (but that should still be confirmed).

For inadvertent re-identification, what is the likelihood that an analyst will know someone in the data set? If the clinical trial was conducted in Japan and the data analyst at the QI is in New York, is there a chance that the QI will know a Japanese participant? The reasonable answer is no, in that inadvertent re-identification will be highly unlikely when the plausibility of a relationship between the participant and the analyst is negligible. Specifically, this means that *Pr*(acquaintance) will be negligibly small. Does that lead us to the conclusion that the data should not be de-identified at all? The answer is no because the Japanese participants will still expect that the data about them are de-identified to some extent. The public perception of the possibility of disclosing data that have a high risk of re-identification needs to be considered.

A **breach** will occur if there is a data breach at the QI's facility. The probability of this type of attack being successful is

$$Pr(\text{re-id, breach}) = Pr(\text{re-id} \mid \text{breach}) \times Pr(\text{breach}) \quad (3)$$

where the term *Pr*(breach) captures the probability that a breach will occur. What should *Pr*(breach) be? Publicly available data about the probability of a breach can be used to determine this value; the value of the conditional in this case, Pr(re-id | breach), will be computed from these data. Data for 2010 show that 19 percent of health care organizations suffered a data breach within the previous year (HIMSS Analytics, 2010); data for 2012 show that this number rose to 27 percent (HIMSS Analytics, 2012). These organizations were all following the HIPAA Security Rule. Note that these figures are averages and may be adjusted to account for variation.

For a nonpublic data release, then, there are three types of attacks for which the re-identification risk needs to be measured and managed. The risk metrics are summarized in Table 6. The overall probability of re-identification will then be the largest value among the three equations.

TABLE 6 Data Risk Metrics

Data Risk	Metric to Use
Pr(re-id | attempt)	Strict average risk
Pr(re-id | acquaintance)	Strict average risk
Pr(re-id | breach)	Strict average risk or maximum risk, depending on the assumptions

Setting Thresholds: What Is Acceptable Risk?

There are quite a few precedents for what can be considered an acceptable amount of risk. These precedents have been in use for many decades, are consistent internationally, and have persisted over time as well (El Emam, 2013). It should be noted, however, that the precedents set to date have been for assessments of maximum risk.

In commentary about the de-identification standard in the HIPAA Privacy Rule, HHS notes in the *Federal Register* (Sweeney, 2002) that

> the two main sources of disclosure risk for de-identified records about individuals are the existence of records with very unique characteristics (e.g., unusual occupation or very high salary or age) and the existence of external sources of records with matching data elements which can be used to link with the de-identified information and identify individuals (e.g., voter registration records or driver's license records) … an expert disclosure analysis would also consider the probability that an individual who is the target of an attempt at re-identification is represented on both files, the probability that the matching variables are recorded identically on the two types of records, the probability that the target individual is unique in the population for the matching variables, and the degree of confidence that a match would correctly identify a unique person.

It is clear that HHS considers unique records to have a high risk of re-identification, but such statements also suggest that nonunique records have an acceptably low risk of re-identification.

Yet uniqueness is not a universal threshold. Historically, data custodians (particularly government agencies focused on reporting statistics) have used the "minimum cell size" rule as a threshold for deciding whether to de-identify data (Alexander and Jabine, 1978; Cancer Care Ontario, 2005; Health Quality Council, 2004a,b; HHS, 2000; Manitoba Center for Health Policy, 2002; Office of the Information and Privacy Commissioner of British Columbia, 1998; Office of the Information and Privacy Commissioner of Ontario, 1994; OMB, 1994; Ontario Ministry of

Health and Long-Term Care, 1984; Statistics Canada, 2007). This rule was originally applied to counting data in tables (e.g., number of males aged 30-35 living in a certain geographic region). The most common minimum cell size in practice is 5, which implies that the maximum probability of re-identifying a record is 1/5, or 0.2. Some custodians, such as certain public health offices, use a smaller minimum count, such as 3 (CDC and HRSA, 2004; de Waal and Willenborg, 1996; NRC, 1993; Office of the Privacy Commissioner of Quebec, 1997; U.S. Department of Education, 2003). Others, by contrast, use a larger minimum, such 11 (in the United States) (Baier et al., 2012; CMS, 2008, 2011; Erdem and Prada, 2011; HHS, 2008a) and 20 (in Canada) (El Emam et al., 2011b, 2012). Based on our review of the literature and the practices of various statistical agencies, the largest minimum cell size is 25 (El Emam et al., 2011b). It should be recognized, however, that there is no agreed-upon threshold, even for what many people would agree is highly sensitive data. For example, minimal counts of 3 and 5 were recommended for HIV/AIDS data (CDC and HRSA, 2004) and abortion data (Statistics Canada, 2007), respectively. Public data releases have used different cell sizes in different jurisdictions. The variability is due, in part, to different tolerances for risk, the sensitivity of data, whether a data sharing agreement is in place, and the nature of the data recipient.

A minimum cell size criterion amounts to a maximum risk value. Yet in some cases, this is too stringent a standard or may not be an appropriate reflection of the type of attack. In such a case, one can use the average risk, as discussed in the previous section. This makes the review of cell size thresholds suitable for both types of risk metrics.

It is possible to construct a decision framework based on these precedents with five "bins" representing five possible thresholds, as shown in Figure 4. At one extreme is data that would be considered identifiable when the cell size is smaller than 3. Next to that are data that

are de-identified with a minimal cell size of 3. Given that this is the least de-identified data set, one could choose to disclose such data sets only to trusted entities where the risks are minimal (for example, where a data sharing agreement is in place, and the data recipient has good security and privacy practices). At the other end of the spectrum is the minimal cell size of 20. This high level of de-identification is appropriate when the data are publicly released, with no restrictions on or tracking of what is done with the data and who has accessed them.

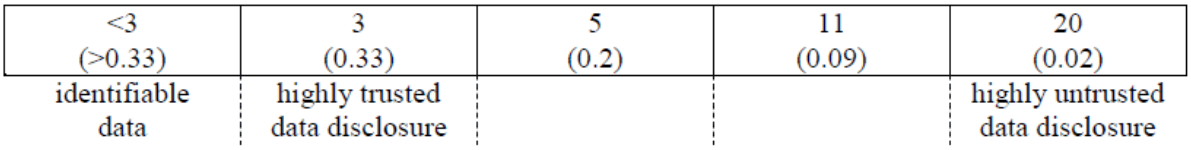

FIGURE 4 Commonly used risk thresholds based on the review/references in the text.

If the extreme situations cannot be justified in a particular disclosure, an alternative process is needed for choosing one of the intermediate values. In Figure 4, this is a choice between a value of 5 and a value of 20.

The above framework does not preclude the use of other values (for example, a sponsor may choose to use a threshold value of 25 observations per cell). However, this framework does ground the choices based on precedents of actual data sets.

What Is the Likelihood of Re-identifying Clinical Trial Data Sets?

There has been concern in the health care and privacy communities that the risk of re-identification in data is quite high and that de-identification is not possible (Ohm, 2010). This argument is often supported by examples of a number of publicly known re-identification attacks. A systematic review of publicly known re-identification attacks found, however, that

when appropriate re-identification standards are used, the risk of re-identification is indeed very small (El Emam et al., 2011a).[4] It was only when no de-identification at all was performed on the data or the de-identification applied was not consistent with or based on best practices that data sets were re-identified with a high success rate. Therefore, the evidence that exists today suggests that using current standards and best practices does provide reasonably strong protections against re-identification.

HOW TO MANAGE RE-IDENTIFICATION RISK

Managing re-identification risk means (1) selecting an appropriate risk metric, (2) selecting an appropriate threshold, and (3) measuring the risk in the actual clinical trial data set that will be disclosed. The choice of a metric is a function of whether the clinical trial data set will be released publicly. For public data sets, it is prudent to use maximum risk in measuring risk and setting thresholds. For nonpublic data sets, a strong case can be made for using average risk (El Emam, 2013; El Emam and Arbuckle, 2013).

How to Choose an Acceptable Threshold

Selecting an acceptable threshold within the range described earlier requires an examination of the context of the data themselves. The re-identification risk threshold is determined based on factors characterizing the QI and the data themselves (El Emam, 2010). These factors have been suggested and have been in use informally by data custodians for at least

[4] Note that this conclusion does not apply to genomic data sets. A discussion of genomic data sets is provided in the last section of this paper.

the last decade and a half (Jabine, 1993a,b). They cover three dimensions (El Emam et al., 2010), as illustrated in Figure 5:

- **Mitigating controls.** This is the set of security and privacy practices that the QI has in place. A recent review identifies a collection of practices used by large data custodians and recommended by funding agencies and IRBs for managing sensitive health information (El Emam et al., 2009).
- **Invasion of privacy.** This entails evaluation of the extent to which a particular disclosure would be an invasion of privacy to the participants (a checklist is available in El Emam et al. [2009]). There are three considerations: (1) the sensitivity of the data (the greater the sensitivity of the data, the greater the invasion of privacy), (2) the potential injury to patients from an inappropriate disclosure (the greater the potential for injury, the greater the invasion of privacy), and (3) the appropriateness of consent for disclosing the data (the less appropriate the consent, the greater the invasion of privacy) (see Box 5).
- **Motives and capacity.** This dimension compasses the motives and the capacity of the QI to re-identify the data, considering such issues as conflicts of interest, the potential for financial gain from re-identification, and whether the data recipient has the skills and financial capacity to re-identify the data (a checklist is available in El Emam et al. [2009]).

In general, many of these elements can be managed through contracts (e.g., a prohibition on re-identification, restrictions on linking the data with other data sets, and disallowing the

sharing of the data with other third parties). For example, if the mitigating controls are low, which means that the QI has poor security and privacy practices, the re-identification threshold should be set at a lower level. This will result in more de-identification being applied. However, if the QI has very good security and privacy practices in place, the threshold can be set higher. Checklists for evaluating these dimensions, as well as a scoring scheme, are available (El Emam, 2013).

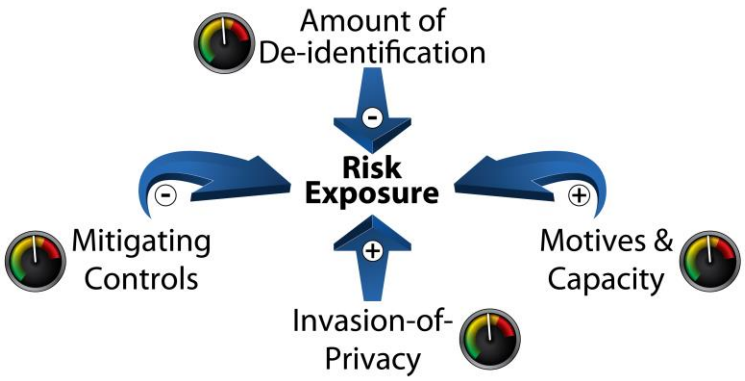

FIGURE 5 Factors to consider when deciding on an acceptable level of re-identification risk. SOURCE: Reprinted with permission from El Eman and colleagues (2014).

BOX 5

Consent and De-identification

As noted earlier, there is no legislative or regulatory requirement to obtain consent from participants to share their de-identified data. There are additional ongoing efforts to ensure that consent forms do not create barriers to data sharing (Health Research Authority, 2013).

Consideration of consent in this context, then, is only to account for situations in which consent has been provided by trial participants or notice has been given to participants. In such cases, the sharing of clinical trial data is not considered as invasive to privacy as opposed to cases in which consent is not sought. Multiple levels of notice and consent can exist for disclosure of de-identified data. These are as follows, in increasing order of invasion of privacy:

- Participants may have consented for the disclosure of personal health data from the trial for the purpose of secondary analysis. This may be a specific or broad consent for secondary analysis. That the sponsor is trying to de-identify the data reflects extra caution and privacy-protective behavior on the part of the sponsor.
- Participants may have consented to the disclosure of only de-identified data for secondary analysis. This may be specific or broad secondary analysis.
- The sponsor does not have express consent for sharing the data, but is consulting with representatives of the trial participants (e.g., patient advocacy groups) and the trial sites to address any sensitivities and to determine the best way to notify participants that their data will be shared.
- The sponsor does not have express consent and is not planning any consultations or

notice.

From a risk management perspective, the first option above is the least invasive of participant privacy, while the last is the most invasive. The practical consequence is that the acceptable threshold (or the definition of "very small risk") will be lower under the most invasive scenario.

If the sponsor is disclosing the data through an online portal, the sponsor has control of many, but not all, of the mitigating controls. This provides additional assurances to the sponsor that a certain subset of controls will be implemented to the sponsor's satisfaction.

Once a threshold has been determined, the actual probability of re-identification is measured in the data set. If the probability is higher than the threshold, transformations of the data need to be performed. Otherwise, the data can be declared to have a very small risk of re-identification.

The implication here is that the amount of data transformation needed will be a function of these other contextual factors. For example, if the QI has good security and privacy practices in place, the threshold chosen will be higher, which means that the data will be subjected to less de-identification.

The security and privacy practices of the QI can be manipulated through contracts. The contract signed by the QI can impose a certain list of practices that must be in place, which are the basis for determining the threshold. Therefore, they must be in place by the QI to justify the level of transformation performed on the data.

This approach is consistent with the limited data set (LDS) method for sharing data under HIPAA. However, this method does not ensure that the risk of re-identification is very small, and therefore the data will still be considered personal health information.

For public data releases, there are no contracts and no expectation that any mitigating controls will be in place. In that case, the lowest probability thresholds (or highest cell size thresholds) are used.

Methods for Transforming the Data

There are a number ways to transform a data set to reduce the probability of re-identification to a value below the threshold. Many algorithms for this purpose have been proposed by the computer science and statistics communities. They vary in quality and performance. Ideally, algorithms adopted for clinical trial data sets should minimize the modifications to the data while ensuring that the measured probability is below the threshold.

Four general classes of techniques have worked well in practice:

- **Generalization.** This is when the value of a field is modified to a more general value. For example, a date of birth can be generalized to a month and year of birth.
- **Suppression.** This is when specific values in the clinical trial data set are removed from the data set (i.e., induced missingness). For example, a value in a record that makes it an outlier may be suppressed.

- **Randomization.** This denotes adding noise to a field. The noise can come from a uniform or other type of distribution. For example, a date may be shifted a week forward or backward.
- **Subsampling.** This is used to disclose a random subset of the data rather than the full data set to the QI.

In practice, a combination of these techniques is applied for any given data disclosure. Furthermore, these techniques can be customized to specific field types. For example, generalization and suppression can be applied differently to dates and zip codes to maximize the data quality for each (El Emam and Arbuckle, 2013).

The application of these techniques can reduce the risk of re-identification. For example, consider the average risk in Table 3, which is 0.59. There is a reduction in average risk to 0.33 when the year of birth is generalized to decades in Table 4. By suppressing some records, it was possible to further reduce the average risk to 0.22 in Table 5. Each transformation progressively reduces the risk.

The Use of Identifier Lists

Thus far we have covered a sufficient number of topics that we can start performing a critical appraisal of some commonly used de-identification methods and the extent to which they can ensure that the risk of re-identification is very small. We focus on the use of identifier lists. The reason is that this approach is quite common, and is being adopted to de-identify clinical trial data.

The HIPAA Privacy Rule's Safe Harbor Standard

We first consider the variable list in the HIPAA Privacy Rule Safe Harbor method.

The Safe Harbor list contains a number of direct identifiers and two quasi-identifiers (i.e., dates and zip codes), as summarized earlier in Box 2. It should be evident that in applying a fixed list of variables, there is no assurance that all of the quasi-identifiers have been accounted for in the risk measurement and the transformation of the data set. For example, other quasi-identifiers, such as race, ethnicity, and occupation, may be in the data set, but they will be ignored. Even if the probability of re-identification under Safe Harbor is small (Benitez and Malin, 2010), this low probability may not carry over with more quasi-identifiers than the two in the original list.

The empirical analysis that was conducted before the Safe Harbor standard was issued assumed that the data set is a random sample from the U.S. population. This assumption may have variable validity in real data sets. However, there will be cases when it is definitely not true. For example, consider a data set that consists of only the records in Table 1. Now, assume that an adversary can find out who is in the data set. This can happen if the data set covers a well-defined population. If the trial site is known, it can be reasonably assumed that the participants in the trial who received treatment at that site live in the same geographic region. If the adversary knows that Bob was born in 1965, lives in the town in which the site is situated, and was in the trial, the adversary knows that Bob is in the data set, and therefore the 27th record must be Bob. This re-identification occurs even though this table meets the requirements of the Safe Harbor standard. Members of a data set may be known if their inclusion in the trial is revealing (e.g., a trial in a workplace where participants have to wear a visible device, parents who must consent

to have their teenage children participate in a study, or adolescents who must miss a few days of school to participate in a study). Therefore, this standard can be protective only if the adversary cannot know who is in the data set. This will be the case if the data set is a random sample from the population.

If these assumptions are met, the applicability of Safe Harbor to a clinical trial data set will be defensible, but only if there are no international participants. If a clinical trial data set includes participants from sites outside the United States, the analysis that justifies using this standard will not be applicable. For example, there is a difference of two orders of magnitude between the median number of individuals living in U.S. zip codes and in Canadian postal codes. Therefore, translating the zip code truncation logic in Safe Harbor to Canadian postal codes would not be based on defensible evidence.

Safe Harbor also has some weaknesses that are specific to the two quasi-identifiers that are included.

In some instances, there may be dates in a clinical trial data set that are not really quasi-identifiers because they do not pass the test highlighted earlier. For example, consider an implantable medical device that fires, and each time it does so there is a time and date stamp in the data stream. The date of a device's firing is unlikely to be a quasi-identifier because it is not knowable, but it is a date.

Safe Harbor states that all three-digit zip codes with fewer than 20,000 inhabitants from the 2010 census must be replaced with "000"; otherwise the three-digit zip code may be included in the data set. The locations of three-digit zip codes with fewer than 20,000 inhabitants are shown in Figure 6. However, in some states there is only one zip code with fewer than 20,000 inhabitants. For example, if a data set is disclosed with "000" for the residential three-digit zip

code for participants in a site in New Hampshire (and it is known that the site is in that state), it is reasonable to assume that the participants also live in that state and to infer that their true three-digit zip code is 036. The same conclusion can be drawn about "000" three-digit zip codes in states such as Alabama, Minnesota, Nebraska, and Nevada.

Other Examples of Identifier Lists

More recent attempts at developing a fixed list of quasi-identifiers to de-identify clinical trial data have indicated that including any combination of two quasi-identifiers (from the prespecified list) is acceptable (Hrynaszkiewicz et al., 2010). Data sets with more than two quasi-identifiers need to go through a more thorough evaluation, such as the risk management approach described earlier. However, this approach suffers from the same limitations as the Safe Harbor standard with respect to the assumption of two quasi-identifiers always having acceptably small risk. An additional limitation is that the authors of the list in Hrynaszkiewicz et al. (2010) present no empirical evaluation demonstrating that this approach consistently produces data sets with a low risk of re-identification, while at least the Safe Harbor list is based on empirical analysis performed by the Census Bureau.

More important, a number of de-identification standards proposed by sponsors have followed similar approaches for sharing clinical trial data from participants globally (see the standards at clinicalstudydatarequest.com). Ideally, methods that can provide stronger assurances should be used to de-identify such data.

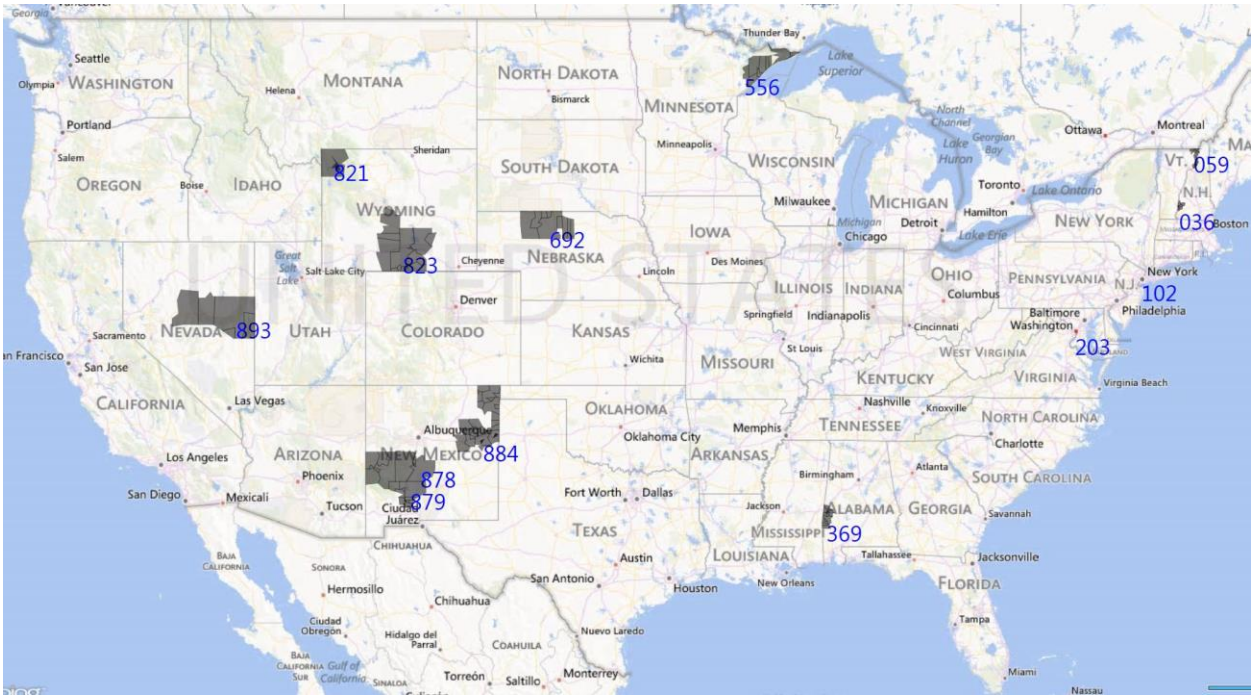

FIGURE 6 Inhabited three-digit zip codes with fewer than 20,000 inhabitants from the 2010 U.S. census.

Putting It All Together

Now that we have gone through the various key elements of the de-identification process, we can put them together into a comprehensive data flow. This flow is illustrated in Figure 7. The steps in this process are as follows.

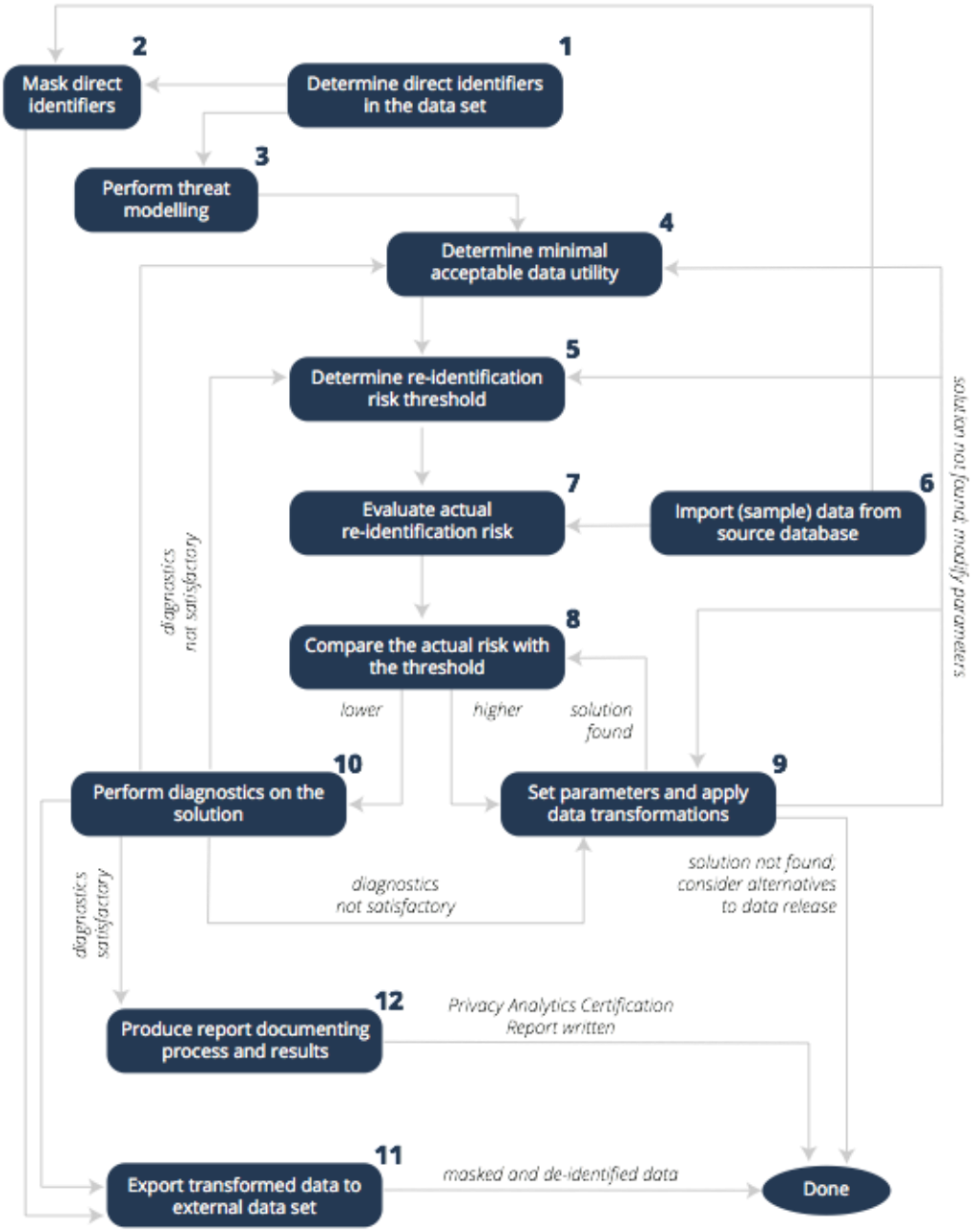

FIGURE 7 The overall de-identification process.
SOURCE: Reprinted with permission from El Emam and colleagues (2014).

Step 1: Determine direct identifiers in the data set.

Determine which fields in the data set are direct identifiers. If the clinical trial data set has already been stripped of direct identifiers, this step may not be necessary.

Step 2: Mask (transform) direct identifiers.

Once the direct identifiers have been determined, masking techniques must be applied to those direct identifiers. Masking techniques include the following: (1) removal of the direct identifiers, (2) replacement of the direct identifiers with random values, or (3) replacement of the direct identifiers with pseudonyms. Once masking has been completed there is virtually no risk of re-identification from direct identifiers. If the database has already been stripped of direct identifiers, this step may not be necessary.

Step 3: Perform threat modeling.

Threat modeling consists of two activities: (1) identification of the plausible adversaries and what information they may be able to access, and (2) determination of the quasi-identifiers in the data set.

Step 4: Determine minimal acceptable data utility.

It is important to determine in advance the minimal relevant data based on the quasi-identifiers. This is essentially an examination of what fields are considered most appropriate given the purpose of the use or disclosure. This step concludes with the imposition of practical limits on how some data may be de-identified and the analyses that may need to be performed later on.

Step 5: Determine the re-identification risk threshold.

This step entails determining what constitutes acceptable risk. As an outcome of the process used to define the threshold, the mitigating controls that need to be imposed on the QI, if any, become evident.

Step 6: Import (sample) data from the source database.

Importing data from the source database may be a simple or complex exercise, depending on the data model of the source data set. This step is included explicitly in the process because it can consume significant resources and must be accounted for in any planning for de-identification.

Step 7: Evaluate the actual re-identification risk.

The actual risk is computed from the data set using the appropriate metric (maximum or strict average). To compute risk, a number of parameters need to be set, such as the sampling fraction.

Step 8: Compare the actual risk with the threshold.

This step entails comparing the actual risk with the threshold determined in Step 5.

Step 9: Set parameters and apply data transformations.

If the measured risk is higher than the threshold, anonymization methods, such as generalization, suppression, randomization, and subsampling, are applied to the data. Sometimes a solution cannot be found within the specified parameters, and it is necessary to go back and reset the parameters. It may also be necessary to modify the threshold and adjust some of the assumptions behind the original risk assessment. Alternatively, some of the assumptions about acceptable data utility may need to be renegotiated with the data users.

Step 10: Perform diagnostics on the solution.

If the measured risk is lower than the threshold, diagnostics should be performed on the solution. Diagnostics may be objective or subjective. An objective diagnostic will evaluate the sensitivity of the solution to violations of assumptions that were made. For example, an assumption may be that an adversary might know the diagnosis code of a patient, or if there is uncertainty about the sampling fraction of the data set, a sensitivity to that value can be performed. A subjective diagnostic will determine whether the utility of the data is sufficiently high for the intended purposes of the use or disclosure.

If the diagnostics are satisfactory, the de-identified data are exported, and a report documenting the de-identification is produced. On the other hand, if the diagnostics are not satisfactory, the re-identification parameters may need to be modified; the risk threshold adjusted; and the original assumptions about minimal, acceptable utility renegotiated with the data user.

Step 11: Export transformed data to external data set.

Exporting the de-identified data to the destination database may be a simple or complex exercise, depending on the data model of the destination database. This step is included explicitly in the process because it can consume significant resources and must be accounted for in any planning for de-identification.

ASSESSING THE IMPACT OF DE-IDENTIFICATION ON DATA QUALITY

As noted above, Safe Harbor and similar methods that significantly restrict the precision of the fields that can be disclosed can result in a nontrivial reduction in the quality of de-

identified data. Therefore, in this section, we focus on data quality when statistical methods are used to de-identify data.

The evidence on the impact of de-identification on data utility is mixed. Some studies show little impact (Kennickell and Lane, 2006), while others show significant impact (Purdam and Elliot, 2007). There is also evidence that data utility will depend on the type of analysis performed (Cox and Kim, 2006; Lechner and Pohlmeier, 2004). In general, if de-identification is accomplished using precise risk measurement and strong optimization algorithms to transform the data, data quality should remain high.

Ensuring that the analysis results produced after de-identification are similar to the results that would be obtained on the original data sets is critical. It would be problematic if a QI attempted to replicate the results from a published trial and were unable to do so because of extensive distortion caused by the de-identification that was applied. Therefore, the amount of distortion must be minimized.

However, de-identification always introduces some distortion, and there is a trade-off between data quality and the amount of de-identification performed to protect privacy. This trade-off can be represented as a curve between data utility and privacy protection as illustrated in Figure 8.

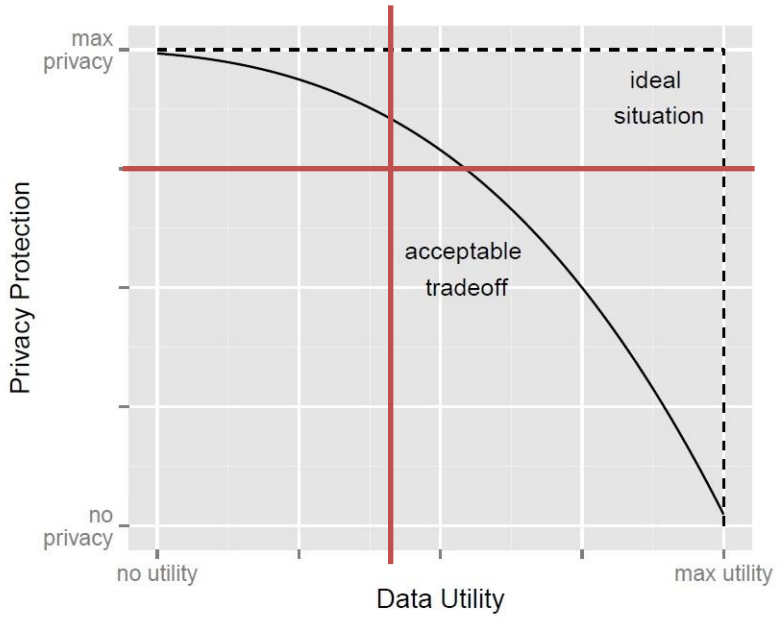

FIGURE 8 The trade-off between privacy and data utility.

Consider, then, that there is a minimal amount of data utility that would be tolerable to ensure that the results of the original trial can be replicated to a large extent. On the other hand, there is a re-identification probability threshold that cannot be exceeded. As shown in Figure 8, this will leave a small range of possible solutions. To ensure that the de-identification solution is truly within this narrow operating range, it is necessary to perform a pilot evaluation on one or more representative clinical trial data sets, and compare the before and after analysis results using exactly the same analytic techniques.

Obtaining similar results for a de-identified clinical trial data set that is intended for public release will be more challenging than disclosing the data set to a QI with strong mitigating controls. The reason is that the amount of de-identification will vary, being more in the former case. This may limit a sponsor's ability to disclose data publicly, or there may have to be a strong replicability caveat on the public data set. For a nonpublic data set when a QI is known, the sponsor may impose a minimal set of mitigating controls through a contract or by providing the

data through an online portal to ensure that the de-identification applied to the data set is not excessive.

GOVERNANCE

Governance is necessary for the sponsor to manage the risks when disclosing clinical trial data, and requires that a set of additional practices be in place. What would be characterized as high-maturity sponsors will have a robust governance process in place.

Governance Practices

Some governance practices are somewhat obvious, such as the need to track all data releases; trigger alerts for data use expirations; and ensure that the documentation for the de-identification for each data release has, in fact, been completed. Other practices are necessary to ensure that participant privacy is adequately protected in practice. Elements of governance practices are listed in Box 6.

BOX 6

Elements of Governance Practices

- Developing and maintaining global anonymization documentation
- Process and tools for tracking all data releases
- Process and tools for triggering alerts for data use expirations

- Ensuring that documentation for the de-identification for each data release is complete and indexed
- On occasion, commissioning controlled re-identification attacks
- Implementing a QI audit process
- Ensuring that there is ethics review that covers protections against attribute disclosure

Controlled Re-identification

The U.K. ICO has recommended that organizations that disclose data also perform controlled re-identification attacks on their disclosed data sets (ICO, 2012). Doing so will allow them to obtain independent evidence on how well their de-identification practices are working and determine whether there are any potential weaknesses that they need to start addressing.

Controlled re-identification attacks are commissioned by the sponsor. With limited funding, these attacks often use publicly available information to attack databases. If additional funding is available, those who conduct these attacks can purchase and use commercial databases to re-identify data subjects.

Appropriate Contracts

Additional governance elements become particularly important when a sponsor discloses data to a QI under a contract. This contract will document the mitigating controls as part of the conditions for receiving the data. The sponsor should then have an audit regime in place to

ensure that QIs have indeed put these practices in place. The sponsor may select high-risk QIs for audit, select randomly, or a combination of the two. Another approach is to ask QIs to conduct third-party audits and report the results back to the sponsor on a regular basis for as long as they are using the data set. The purpose of the audit is to ensure that the mitigating controls are indeed in place.

Enterprise De-identification Process

At an enterprise level, sponsors need to have an enterprise de-identification process that will be applied across all clinical trial data sets. This process includes the appropriate thresholds and controls for data releases, as well as templates for data sharing agreements and terms of use of data. The global process ensures consistency across all data releases. This process must then be enacted for each clinical trial data set, and this may involve some customization to address specific characteristics of a given data set.

The cost of such a process will depend on the size of the sponsor and the heterogeneity of its clinical trials and therapeutic areas. However, in the long term such an approach can be expected to have a lower total cost since there will be more opportunities for reuse and learning.

In practice, many sponsors have standard case report forms (CRFs) for a subset of the data they collect in their clinical trials. For example, there may be standard CRFs for demographics or for standardized measures and patient-reported outcomes. The global process can classify the variables in these standard CRFs as direct and quasi-identifiers and articulate the techniques that should be used to transform those variables. This will reduce the anonymization effort per clinical trial by a nontrivial amount.

Protecting Against Attribute Disclosure

At the beginning of this paper, we briefly mentioned attribute disclosure, but did not address how to protect against it. Such protections can be implemented as part of governance. However, in general, modifying the data to protect against attribute disclosure means reducing the plausible inferences that can be drawn from the data. This can be detrimental to the objective of learning as much as possible from the data and building generalizable statistical models from the data. Furthermore, to protect against attribute disclosure, one must anticipate all inferences and make data modifications to impede them, which may not be possible.

Some inferences may be desirable because they may enhance understanding of the treatment benefits or safety of a new drug or device, and some inferences will be stigmatizing to the data subjects. One will not want to make modifications to the data that block the former type of inferences.

For nonpublic data releases, it is recommended that there be an ethics review of the analysis protocols. As part of the ethics review process, the ethics committee or council will examine the potential for stigmatizing attribute disclosure. This is a subjective decision and will have to take into account current social norms and participant expectations (see also the discussion in El Emam and Arbuckle [2013]). The ethics review may be performed on the secondary analysis protocol by the QI's institutional IRB, or by a separate committee reporting to the sponsor or even within the sponsor. Such an approach will maximize data integrity but also provide assurance that attribute disclosure is addressed. An internal sponsor ethics review council will include a privacy professional, an ethicist, a lay person representing the participants,

a person with knowledge of the clinical trials business at the sponsor, and a brand or public relations person.

For public data releases, there is no analysis protocol or a priori approval process, and therefore it will be challenging to provide assurances about attribute disclosure.

De-identifying Genomic Data

There have been various proposals to apply the types of generalization and randomization strategies discussed in this paper to genomic data, and *omics data more generally (e.g., RNA expression or proteomic records) (Li et al., 2012; Lin et al., 2002, 2004; Malin, 2005). However, evidence suggests that such methods may not be suitable for the anonymization of biomarkers that constitute a large number of dimensions. The main reasons are that they can cause significant distortion of long sequences, and the assumptions that need to be made to de-identify sequences of patient events (e.g., visits and claims) will not apply to *omic data. At the same time, there are nuances that are worth considering. For context, we address concerns around genomic data specifically, while noting that similar allusions can be made to other types of data.

First, it is important to recognize that many of the attacks that have been carried out on genomic data require additional information (Malin et al., 2011). In certain cases, for instance, the re-identification of genomic data is accomplished through the demographics of the corresponding research participant; the associated clinical information (Loukides et al., 2010b); or contextual cues associated with the collection and dissemination of the data, such as the set of health care providers visited by the participant (Malin and Sweeney, 2004). For example, a recently reported re-identification attack on participants in the Personal Genome Project (PGP)

was based almost entirely on information derived from publicly accessible profiles—notably birth date (or month and year), gender, and geographic indicators of residence (e.g., zip code) (Sweeney et al., 2013). Other individuals in the PGP were re-identified based on the fact that they uploaded compressed files that incorporated their personal names as file names when uncompressed. This attack used the same type of variables that can be protected using the techniques described in this paper. Moreover, it has been shown that many of the protection strategies discussed in this paper can be tailored to support genome-phenome association discovery (e.g., through anonymization of standardized clinical codes [Heatherly et al., 2013; Loukides et al., 2010a]).

This fact is true for attacks that factor genomic data into the attack as well. For instance, it was recently shown that an adversary could use publicly available databases that report on Y-chromosme–surname correlations to ascertain the surname of a genome sequence lacking an individual's name (Haggie, 2013). However, for this attack to be successful, it required additional information about the corresponding individual. Specifically, the attacker also needed to know the approximate area of residence (e.g., U.S. state) and approximate age of the individual. While such information may be permitted within a Safe Harbor de-identification framework, a statistical assessment of the potential identifiability of such information would indicate that such ancillary information might constitute an unacceptably high rate of re-identification risk. At the same time, it should be recognized that, even when such information was made available, the attack reported in Haggie (2013) was successful 12 percent of the time and unsuccessful 5 percent of the time. In other words, there is variability in the chance that such attacks will be successful.

More direct attacks are, however, plausible. There is evidence that a sequence of 30 to 80 independent single nucleotide polymorphisms (SNPs) could uniquely identify a single person (Lin et al., 2004). Unlike the surname inference attack mentioned above, a direct attack would require that the adversary already have identified genotype data for a target individual. Yet linking an individual using his or her genome would permit the adversary to learn any additional information in the new resource, such as the individual's health status. Additionally, a recent demonstration with data from openSNP and Facebook suggests that in certain instances, the genomic status of an individual can be inferred based on the genome sequences of close family members (Humbert et al., 2013).

Beyond direct matching of sequences, there is also a risk of privacy compromise in "pooled" data, where only summary statistics are reported. For instance, it has been shown that it is possible to determine whether an individual is in a pool of cases or controls for a study by assessing the likelihood that the individual's sequence is "closer" to one group or the other (Homer et al., 2008; Jacobs et al., 2009; Wang et al., 2009). Despite such vulnerability, it has also been shown that the likelihood of success for this attack becomes lower as the number of people in each group increases. In fact, for studies with a reasonable number of participants (more than 1,000), it is safe to reveal the summary statistics of all common (not rare) genomic regions (Sankararaman et al., 2009).

However, one of the challenges with genomic data is that it is possible to learn phenotypic information directly. When such information can be ascertained with certainty, it can then be used in a re-identification attack. For example, predictions (varying in accuracy) of height, facial morphology, age, body mass index, approximate skin pigmentation, eye color, and diagnosis of cystic fibrosis or Huntington's chorea from genetic information have been reported

(Kayser and de Knijff, 2011; Kohn, 1991; Lowrance and Collins, 2007; Malin and Sweeney, 2000; Ou et al., 2012; Silventoinen et al., 2003; Wjst, 2010; Zubakov et al., 2010), although it should be noted that there have been no full demonstrations of attacks using such inferences. Also, because of the errors in some of these predictions (excluding Mendelian disorders that are directly dependent on a mutation in a certain portion of the genome), it is not clear that they would be sufficiently reliable for re-identification attacks.

Although traditional generalization and randomization strategies may not provide a sufficient balance between utility and privacy for high-dimensional *omics data, a solution to the problem may be possible with the assistance of modern cryptography. In particular, secure multiparty computation (SMC) corresponds to a set of techniques (and protocols) that allow quite sophisticated mathematical and statistical operations to be performed on encrypted data. In the process, individual records would never be disclosed to the user of such a resource. This type of protection would not prevent inference through summary-level statistics, but it would prevent direct attacks on individuals' records. SMC solutions have been demonstrated that have been tailored to support frequency queries (Kantarcioglu et al., 2008), genomic sequence alignment (Chen et al., 2012), kinship (and other comparison) tests (Baldi et al., 2011; He et al., 2014) and personalized medical risk scores (Ayday et al., 2013a,b). Nonetheless, the application of these methods to genetic data is still in the early stages of research, and it may be a few more years before some large-scale practical results are seen.

REFERENCES

Alexander, L., and T. Jabine. 1978. Access to social security microdata files for research and statistical purposes. *Social Security Bulletin* 41(8):3-17.

Article 29 Data Protection Working Party. 2007. *Opinion 4/2007 on the concept of personal data*. WP136. http://ec.europa.eu/justice/policies/privacy/docs/wpdocs/2007/wp136_en.pdf (accessed December 19, 2014).

Article 29 Data Protection Working Party. 2014. *Opinion 05/2014 on anonymization techniques*. WP216. http://ec.europa.eu/justice/data-protection/article-29/documentation/opinion-recommendation/files/2014/wp216_en.pdf (accessed December 19, 2014).

Ayday, E., J. L. Raisaro, and J.-P. Hubaux. 2013a. Privacy-enhancing technologies for medical tests using genomic data. *20th Annual Network and Distributed System Security Symposium (NDSS)*.

Ayday, E., J. L. Raisaro, J.-P. Hubaux, and J. Rougemont. 2013b. Protecting and evaluating genomic privacy in medical tests and personalized medicine. *Proceedings of the 12th ACM Workshop on Workshop on Privacy in the Electronic Society* 95-106.

Baier, P., S. Hinkins, and F. Scheuren. 2012. *The electronic health records incentive program eligible professionals public use file*. http://www.cms.gov/Regulations-and-Guidance/Legislation/EHRIncentivePrograms/DataAndReports.html (accessed December 19, 2014).

Baldi, P., R. Baronio, E. De Cristofaro, P. Gasti, and G. Tsudik. 2011. Countering GATTACA: Efficient and secure testing of fully-sequenced human genomes. *Proceedings of the 18th ACM Conference on Computer and Communications Security* 691-702.

Benitez, K., and B. Malin. 2010. Evaluating re-identification risks with respect to the HIPAA privacy rule. *Journal of the American Medical Informatics Association* 17(2):169-177.

Bhattacharjee, Y. 2012. Pharma firms push for sharing of cancer trial data. *Science* 38(6103):29.

Canadian Institute for Health Information. 2010. *'Best practice' guidelines for managing the disclosure of de-identified health information*. http://www.ijpc-se.org/documents/hhs10.pdf (accessed December 19, 2014).

Cancer Care Ontario. 2005. *Cancer Care Ontario data use and disclosure policy*. Toronto, ON: Cancer Care Ontario.

Castellani, J. 2013. Are clinical trial data shared sufficiently today? Yes. *British Medical Journal* 347(1):f1881.

CDC (Centers for Disease Control and Prevention) and HRSA (Health Resources and Services Administration). 2004. *Integrated guidelines for developing epidemiologic profiles: HIV Prevention and Ryan White CARE Act community planning*. Atlanta, GA: CDC. http://www.cdph.ca.gov/programs/aids/Documents/GLines-IntegratedEpiProfiles.pdf (accessed December 19, 2014).

Center for Business and Information Technologies. 2013. *Cajun Code Fest*. http://cajuncodefest.org/ (accessed November 9, 2012).

Chen, Y., B. Peng, X. Wang, and H. Tang. 2012. Large-scale privacy-preserving mapping of human genomic sequences on hybrid clouds. *Proceeding of the 19th Network and Distributed System Security Symposium*.

CMS (Centers for Medicare & Medicaid Services). 2008. *2008 basic stand alone Medicare claims public use files*. http://www.cms.gov/Research-Statistics-Data-and-

Systems/Statistics-Trends-and-Reports/BSAPUFS/Downloads/2008_BSA_PUF_Disclaimer.pdf (accessed December 19, 2014).
CMS. 2011. *BSA inpatient claims PUF*. http://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/BSAPUFS/Inpatient_Claims.html (accessed December 19, 2014).
Cohen, J. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* XX(1):37-46.
CORE (Center for Outcomes Research and Evaluation). 2014. *The YODA Project*. http://medicine.yale.edu/core/projects/yodap/ (accessed December 19, 2014).
Cox, L. H., and J. J. Kim. 2006. Effects of rounding on the quality and confidentiality of statistical data. In *Privacy in statistical databases*, edited by J. Domingo-Ferrer and L. Franconi. New York: Springer-Verlag Berlin Heidelberg. Pp. 48-56.
Dankar, F., K. E. Emam, A. Neisa, and T. Roffey. 2012. Estimating the re-identification risk of clinical data sets. *BMC Medical Informatics and Decision Making* 12(1):66.
de Waal, A., and L. Willenborg. 1996. A view on statistical disclosure control for microdata. *Survey Methodology* 22(1):95-103.
Dryad. undated. *Dryad digital repository*. http://datadryad.org/ (accessed September 19, 2013).
El Emam, K. 2010. Risk-based deidentification of health data. *IEEE Security and Privacy* 8(3):64-67.
El Emam, K. 2013. *Guide to the deidentification of personal health information*. Boca Raton, FL: CRC Press (Auerbach Publications).
El Emam, K., and L. Arbuckle. 2013. *Anonymizing health data: Case studies and methods to get you started*. Sebastopol, CA: O'Reilly Media.
El Emam, K., F. Dankar, R. Vaillancourt, T. Roffey, and M. Lysyk. 2009. Evaluating patient re-identification risk from hospital prescription records. *Canadian Journal of Hospital Pharmacy* 62(4):307-319.
El Emam, K., A. Brown, P. AbdelMalik, A. Neisa, M. Walker, J. Bottomley, and T. Roffey. 2010. A method for managing re-identification risk from small geographic areas in Canada. *BMC Medical Informatics and Decision Making* 10(1):18.
El Emam, K., E. Jonker, L. Arbuckle, and B. Malin, 2011a. A systematic review of re-identification attacks on health data. *PLOS ONE* 6(12):e28071.
El Emam, K., D. Paton, F. Dankar, and G. Koru. 2011b. Deidentifying a public use microdata file from the Canadian national discharge abstract database. *BMC Medical Informatics and Decision Making* 11:53.
El Emam, K., L. Arbuckle, G. Koru, B. Eze, L. Gaudette, E. Neri, S. Rose, J. Howard, and J. Gluck. 2012. Deidentification methods for open health data: The case of the heritage health prize claims dataset. *Journal of Medical Internet Research* 14(1):e33.
El Emam, K., G. Middleton, and L. Arbuckle. 2014. *An implementation guide for data anonymization*. Bloomington, IN: Trafford Publishing.
EMA (European Medicines Agency). 2014a. *European Medicines Agency policy on publication of clinical data for medicinal products for human use*. http://www.ema.europa.eu/docs/en_GB/document_library/Other/2014/10/WC500174796.pdf (accessed December 19, 2014).

EMA. 2014b. *Release of data from clinical trials*. http://www.ema.europa.eu/ema/index.jsp?curl=pages/special_topics/general/general_content_000555.jsp&mid=WC0b01ac0580607bfa (accessed December 19, 2014).
Erdem, E., and S. I. Prada. 2011. *Creation of public use files: Lessons learned from the comparative effectiveness research public use files data pilot project*. http://mpra.ub.uni-muenchen.de/35478/ (accessed November 9, 2012).
Fung, B. C. M., K. Wang, R. Chen, and P. S. Yu. 2010. Privacy-preserving data publishing: A survey of recent developments. *ACM Computing Surveys* 42(4):1-53.
Gøtzsche, P. C. 2011. Why we need easy access to all data from all clinical trials and how to accomplish it. *Trials* 12(1):249.
Haggie, E. 2013. *PLOS Genetics partners with Dryad*. http://blogs.plos.org/biologue/2013/09/18/plos-genetics-partners-with-dryad/ (accessed September 19, 2013).
Harrison, C. 2012. GlaxoSmithKline opens the door on clinical data sharing. *Nature Reviews Drug Discovery* 11(12):891-892.
He, D., N. A. Furlotte, F. Hormozdiari, J. W. J. Joo, A. Wadia, R. Ostrovsky, A. Sahai, and E. Eskin. 2014. Identifying genetic relatives without compromising privacy. *Genome Research* 24(4):664-672.
Health Quality Council. 2004a. *Privacy code*. Saskatoon, Canada: Health Quality Council.
Health Quality Council. 2004b. *Security and confidentiality policies and procedures*. Saskatoon, Canada: Health Quality Council.
Health Research Authority. 2013. *The HRA interest in good research conduct: Transparent research*. http://www.hra.nhs.uk/documents/2013/08/transparent-research-report.pdf (accessed December 19, 2014).
Heatherly, R. D., G. Loukides, J. C. Denny, J. L. Haines, D. M. Roden, and B. A. Malin. 2013. Enabling genomic-phenomic association discovery without sacrificing anonymity. *PLOS ONE* 8(2):e53875.
Hede, K. 2013. Project data sphere to make cancer clinical trial data publicly available. *Journal of the National Cancer Institute* 105(16):1159-1160).
HHS (U.S. Department of Health and Human Services). 2000. *Standards for privacy of individually identifiable health information*. Washington, DC: HHS.
HHS. 2004. *Guidance on research involving coded private information or biological specimens*. Washington, DC: HHS.
HHS. 2008a. *Instructions for Completing the Limited Data Set Data Use Agreement (DUA) (CMS-R-0235L)*. http://innovation.cms.gov/Files/x/Bundled-Payments-for-Care-Improvement-Data-Use-Agreement.pdf (accessed December 19, 2014).
HHS. 2008b. *OHRP: Guidance on research involving coded private information or biological specimens*. Washington, DC: HHS.
HHS. 2012. *Guidance regarding methods for deidentification of protected health information in accordance with the Health Insurance Portability and Accountability Act (HIPAA) privacy rule*. Washington, DC: HHS.
HIMSS Analytics. 2010. 2010 *HIMSS Analytics report: Security of patient data*. Chicago, IL: HIMSS Analytics.
HIMSS Analytics. 2012. *2012 HIMSS Analytics report: Security of patient data*. Chicago, IL: HIMSS Analytics.

Homer, N., S. Szelinger, M. Redman, D. Duggan, W. Tembe, J. Muehling, J. Pearson, D. Stephan, S. Nelson, and D. Craig. 2008. Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. *PLOS Genetics* 4(8):e1000167.

Hrynaszkiewicz, I., M. L. Norton, A. J. Vickers, and D. G. Altman. 2010. Preparing raw clinical data for publication: Guidance for journal editors, authors, and peer reviewers. *Trials* 11(1):9.

Humbert, M., E. Ayday, J.-P. Hubaux, and A. Telenti. 2013. Addressing the concerns of the Lacks family: Quantification of kin genomic privacy. *Proceedings of the 2013 ACM SIGSAC Conference on Computer and Communications Security* 1141-1152.

ICO (Information Commissioner's Office). 2012. *Anonymisation: Managing data protection risk code of practice*. http://ico.org.uk/~/media/documents/library/Data_Protection/Practical_application/anonymisation-codev2.pdf (accessed December 19, 2014).

ImmPort. undated. *ImmPort: Immunology database and analysis portal*. https://immport.niaid.nih.gov/immportWeb/home/home.do?loginType=full (accessed September 19, 2013).

IOM (Institute of Medicine). 2013. *Sharing clinical research data: Workshop summary*. Washington, DC: The National Academies Press.

ISO (International Organization for Standardization). 2008. *Health informatics—Pseudonymization*. ISO/TS 25237:2008. Geneva, Switzerland: ISO.

Jabine, T. 1993a. Procedures for restricted data access. *Journal of Official Statistics* 9(2):537-589.

Jabine, T. 1993b. Statistical disclosure limitation practices of United States statistical agencies. *Journal of Official Statistics* 9(2):427-454.

Jacobs, K. B., M. Yeager, S. Wacholder, D. Craig, P. Kraft, D. J. Hunter, J. Paschal, T. A. Manolio, M. Tucker, R. N. Hoover, G. D. Thomas, S. J. Chanock, and N. Chatterjee. 2009. A new statistic and its power to infer membership in a genome-wide association study using genotype frequencies. *Nature Genetics* 41(11):1253-1257.

Kantarcioglu, M., W. Jiang, Y. Liu, and B. Malin. 2008. A cryptographic approach to securely share and query genomic sequences. *IEEE Transactions on Information Technology in Biomedicine* 12(5):606-617.

Kayser, M., and P. de Knijff. 2011. Improving human forensics through advances in genetics, genomics and molecular biology. *Nature Reviews Genetics* 12(3):179-192.

Kennickell, A., and J. Lane. 2006. Measuring the impact of data protection techniques on data utility: Evidence from the survey of consumer finances. In *Privacy in statistical databases*, edited by J. Domingo-Ferrer and L. Franconi. New York: Springer-Verlag Berlin Heidelberg. Pp. 291-303.

Knoppers, B. M., and M. Saginur. 2005. The Babel of genetic data terminology. *Nature Biotechnology* 23(8):925-927.

Kohn, L. A. P. 1991. The role of genetics in craniofacial morphology and growth. *Annual Review of Anthropology* 20(1):261-278.

Krumholz, H. M., and J. S. Ross. 2011. A model for dissemination and independent analysis of industry data. *Journal of the American Medical Association* 306(14):1593-1594.

Lechner, S., and W. Pohlmeier. 2004. To blank or not to blank? A comparison of the effects of disclosure limitation methods on nonlinear regression estimates. In *Privacy in statistical*

databases, edited by J. Domingo-Ferrer and V. Torra. New York: Springer-Verlag Berlin Heidelberg. Pp. 187-200.
Li, G., Y. Wang, and X. Su. 2012. Improvements on a privacy-protection algorithm for DNA sequences with generalization lattices. *Computer Methods and Programs in Biomedicine* 108(1):1-9.
Lin, Z., M. Hewett, and R. Altman. 2002. Using binning to maintain confidentiality of medical data. *Proceedings of the American Medical Informatics Association Annual Symposium* 454-458.
Lin, Z., A. Owen, and R. Altman. 2004. Genomic research and human subject privacy. *Science* 305:183.
Loukides, G., A. Gkoulalas-Divanis, and B. Malin. 2010a. Anonymization of electronic medical records for validating genome-wide association studies. *Proceedings of the National Academy of Sciences* 107(17):7898-7903.
Loukides, G., J. C. Denny, and B. Malin. 2010b. The disclosure of diagnosis codes can breach research participants' privacy. *Journal of the American Medical Informatics Association* 17(3):322-327.
Lowrance, W., and F. Collins. 2007. Identifiability in genomic research. *Science* 317:600-602.
Machanavajjhala, A., J. Gehrke, and D. Kifer. 2007. l-Diversity: Privacy beyond k-anonymity. *Transactions on Knowledge Discovery from Data* 1(1):1-47.
Malin, B. 2005. Protecting genomic sequence anonymity with generalization lattices. *Methods of Information in Medicine* 44:687-692.
Malin, B., and L. Sweeney. 2000. Determining the identifiability of DNA database entries. *Proceedings of the AMIA Symposium* 537-541.
Malin, B., and L. Sweeney. 2004. How (not) to protect genomic data privacy in a distributed network: Using trail re-identification to evaluate and design anonymity protection systems. *Journal of Biomedical Informatics* 37(3):179-192.
Malin, B., G. Loukides, K. Benitez, and E. W. Clayton. 2011. Identifiability in biobanks: Models, measures, and mitigation strategies. *Human Genetics* 130(3):383-392.
Manitoba Center for Health Policy. 2002. *Privacy code*. http://umanitoba.ca/faculties/medicine/units/mchp/media_room/media/MCHP_privacy_code.pdf (accessed December 19, 2014).
Mello, M. M. J. K. Francer, M. Wilenzick, P. Teden, B. E. Bierer, and M. Barnes. 2013. Preparing for responsible sharing of clinical trial data. *New England Journal of Medicine* 369(17):1651-1658.
MRC (Medical Research Council). 2011. *Data sharing*. http://www.mrc.ac.uk/research/research-policy- ethics/data-sharing/ (accessed December 19, 2014).
NIH (National Institutes of Health). 2003. *Final NIH statement on sharing research data*. http://grants.nih.gov/grants/guide/notice-files/NOT-OD-03-032.html (accessed December 19, 2014).
Nisen, P., and F. Rockhold, 2013. Access to patient-level data from GlaxoSmithKline clinical trials. *New England Journal of Medicine* 369(5):475-478.
NRC (National Research Council). 1993. *Private lives and public policies: Confidentiality and accessibility of government statistics*. Washington, DC: National Academy Press.
Office of the Information and Privacy Commissioner of British Columbia. 1998. *Order No. 261-1998*. https://www.oipc.bc.ca/orders/496 (accessed December 19, 2014).

Office of the Information and Privacy Commissioner of Ontario. 1994. *Order P-644*. http://www.ipc.on.ca/images/Findings/Attached_PDF/P-644.pdf (accessed December 19, 2014).
Office of the Privacy Commissioner of Quebec (CAI). 1997. *Chenard v. Ministere de l'agriculture, des pecheries et de l'alimentation* (141).
Ohm, P. 2010. Broken promises of privacy: Responding to the surprising failure of anonymization. *UCLA Law Review* 57:1701.
OMB (Office of Management and Budget). 1994. *Report on statistical disclosure limitation methodology*. Working paper 22. Washington, DC: OMB.
Ontario Ministry of Health and Long-Term Care. 1984. *Corporate policy 3-1-21*. Toronto, ON: Ontario Ministry of Health and Long-Term Care.
Ou, X., J. Gao, H. Wang, H. Wang, H. Lu, and H. Sun. 2012. Predicting human age with bloodstains by sjTREC quantification. *PLOS ONE* 7(8):e42412.
Perun, H., M. Orr, and F. Dimitriadis. 2005. *Guide to the Ontario Personal Health Information Protection Act*. Toronto, ON: Irwin Law.
Purdam, K., and M. Elliot. 2007. A case study of the impact of statistical disclosure control on data quality in the individual UK samples of anonymised records. *Environment and Planning A* 39(5):1101-1118.
Rothstein, M. 2005. Research privacy under HIPAA and the Common Rule. *Journal of Law, Medicine & Ethics* 33:154-159.
Rothstein, M. 2010. Is deidentification sufficient to protect health privacy in research. *The American Journal of Bioethics* 10(9):3-11.
Samarati, P. 2001. Protecting respondents' identities in microdata release. *IEEE Transactions on Knowledge and Data Engineering* 13(6):1010-1027.
Sandercock, P. A., M. Niewada, A. Członkowska, and the International Stroke Trial Collaborative Group. 2011. The International Stroke Trial database. *Trials* 12(1):101.
Sankararaman, S., G. Obozinski, M. I. Jordan, and E. Halperin. 2009. Genomic privacy and limits of individual detection in a pool. *Nature Genetics* 41(9):965-967.
Silventoinen, K., S. Sammalisto, M. Perola, D. I. Boomsma, B. K. Cornes, C. Davis, L. Dunkel, M. De Lange, J. R. Harris, J. V. B. Hjelmborg, M. Luciano, N. G. Martin, J. Mortensen, L. Nisticò, N. L. Pedersen, A. Skytthe, T. D. Spector, M. A. Stazi, G. Willemsen, and J. Kaprio. 2003. Heritability of adult body height: A comparative study of twin cohorts in eight countries. *Twin Research* 6(5):399-408.
Skinner, C. J. 1992. On identification disclosure and prediction disclosure for microdata. *Statistica Neerlandica* 46(1):21-32.
Skinner, C., and N. Shlomo. 2008. Assessing identification risk in survey microdata using log-linear models. *Journal of the American Statistical Association* 103(483):989-1001.
Statistics Canada. 2007. *Therapeutic abortion survey*. http://www23.statcan.gc.ca/imdb/p2SV.pl?Function=getSurvey&SurvId=1062&InstaId=31176&SDDS=3209 (accessed December 19, 2014).
Sweeney, L. 2002. k-anonymity: A model for protecting privacy. *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems* 10(5):557-570.
Sweeney, L., A. Abu, and J. Winn. 2013. *Identifying participants in the personal genome project by name*. http://dataprivacylab.org/projects/pgp/1021-1.pdf (accessed December 19, 2014).

U.S. Department of Education. 2003. *NCES statistical standards*. http://nces.ed.gov/pubs2003/2003601.pdf (accessed December 19, 2014).

Vallance, P., and I. Chalmers. 2013. Secure use of individual patient data from clinical trials. *The Lancet* 382(9898):1073-1074.

Wang, R., Y. F. Li, X. Wang, H. Tang, and X. Zhou. 2009. *Learning your identity and disease from research papers: Information leaks in genome wide association study*. http://www.informatics.indiana.edu/xw7/papers/gwas_paper.pdf (accessed December 19, 2014).

Wellcome Trust. 2011. *Sharing research data to improve public health: Full joint statement by funders of health research*. http://www.wellcome.ac.uk/About-us/Policy/Spotlight-issues/Data-sharing/Public-health-and-epidemiology/WTDV030690.htm (accessed December 19, 2014).

Willenborg, L., and T. de Waal. 1996. *Statistical disclosure control in practice*. New York: Springer-Verlag.

Willenborg, L., and T. de Waal. 2001. *Elements of statistical disclosure control*. New York: Springer-Verlag.

Wjst, M. 2010. Caught you: Threats to confidentiality due to the public release of large-scale genetic data sets. *BMC Medical Ethics* 11:21.

Zubakov, D., F. Liu, M. C. van Zelm, J. Vermeulen, B. A. Oostra, C. M. van Duijn, G. J. Driessen, J. J. M. van Dongen, M. Kayser, and A. W. Langerak. 2010. Estimating human age from T-cell DNA rearrangements. *Current Biology* 20(22):R970-R971.