Avoiding disclosure of individually identifiable health information: a literature review

Sergio I Prada and Claudia Gonzalez and Joshua Borton and Johannes Fernandes-Huessy and Craig Holden and Elizabeth Hair and Tim Mulcahy

IMPAQ International, Optimal Solutions Group, NORC at the University of Chicago

14. December 2011

Avoiding Disclosure of Individually Identifiable Health Information : A Literature Review

Sergio I. Prada, Claudia González-Martínez, Joshua Borton, Johannes Fernandes-Huessy, Craig Holden, Elizabeth Hair and and Tim Mulcahy

SAGE Open published online 14 December 2011
DOI: 10.1177/2158244011431279

SAGE Open
1–16
© The Author(s) 2011
DOI: 10.1177/2158244011431279
http://sgo.sagepub.com

Avoiding Disclosure of Individually Identifiable Health Information: A Literature Review

Sergio I. Prada[1], Claudia González-Martínez[2], Joshua Borton[3], Johannes Fernandes-Huessy[3], Craig Holden[3], Elizabeth Hair[3], and Tim Mulcahy[3]

Abstract

Achieving data and information dissemination without harming anyone is a central task of any entity in charge of collecting data. In this article, the authors examine the literature on data and statistical confidentiality. Rather than comparing the theoretical properties of specific methods, they emphasize the main themes that emerge from the ongoing discussion among scientists regarding how best to achieve the appropriate balance between data protection, data utility, and data dissemination. They cover the literature on de-identification and reidentification methods with emphasis on health care data. The authors also discuss the benefits and limitations for the most common access methods. Although there is abundant theoretical and empirical research, their review reveals lack of consensus on fundamental questions for empirical practice: How to assess disclosure risk, how to choose among disclosure methods, how to assess reidentification risk, and how to measure utility loss.

Keywords

public use files, disclosure avoidance, reidentification, de-identification, data utility

Introduction

The U.S. Government is devoting unprecedented attention to the health care sector. Among the different initiatives is the policy of increasing the openness of information by providing the public with better access to federal data sets. Achieving data and information dissemination without harming anyone is a central task of any entity in charge of collecting data. The balance lies in protecting the privacy of those in the data while minimizing data utility loss (Kinney, Karr, & Gonzalez, 2009). Although the need for such balance is true of every data set, it becomes more critical when the information collected is about personal health status and personal health care received.

The benefits of data dissemination are potentially enormous. Rigorous research providing information about the quality, efficiency, effectiveness, and safety of the health care received by members of society can inform and guide decisions in all public policy debates. At the same time, data privacy and confidentiality violations could be potentially disastrous for individuals or groups (Rothstein, 2010). Such a breach, in turn, could cause irreparable damage to the credibility of the data collector and disseminator.

In this article, we examine the literature on data and statistical confidentiality. Rather than comparing the theoretical properties of specific methods, we emphasize the main themes that emerge from the ongoing discussion among scientists regarding how best to achieve the appropriate balance between data protection and data dissemination. With that objective, we examine the literature published in academic journals and books and proceedings from specialized conferences. The fields in which much of the discussion is concentrated include statistics, computer science, data privacy and security, electrical engineering, bioinformatics, and health services.

This article provides a summary of key concepts and issues in the literature. It is designed to be a point of entry for policy makers, researchers, and practitioners in the health care sector who are new to the literature and for those considering making data sets publicly available. The article discusses only the literature on statistical disclosure methods. It does not discuss computational disclosure control (i.e., computer programs that maintain anonymity by automatically generalizing, substituting, and removing information when users access the data), methods for tabular data, attribute disclosure (i.e., disclosure of sensitive information other than direct identifiers), or

[1]IMPAQ International LLC, Columbia, MD, USA
[2]Optimal Solutions Group LLC, College Park, MD, USA
[3]NORC at the University of Chicago, Bethesda, MD, USA

Corresponding Author:
Sergio Prada, IMPAQ International LLC, 10420 Little Patuxent Parkway, Suite 300, Columbia, MD 21044, USA
Email: sprada@impaqint.com

inferential disclosure (i.e., information that can be inferred about a record in a data set with better accuracy). There is significant literature on each of these topics, which are beyond the scope of this article.

Our article is divided into six sections, of which this "Introduction" is the first. The second section presents "The Policy and Academic Context" surrounding the discussion. The third section discusses the state of the art in "De-Identification Methods," while the fourth emphasizes the state of the art in "Reidentification Methods." The fifth section presents the conclusions from the literature on the different ways in which users may "Access" public data, stressing the trade-offs between (a) confidentiality and utility and (b) confidentiality and ease of access. The last section presents the "Conclusion."

The Policy and Academic Context

Historic Perspective

Concerns about privacy and confidentiality in governmental efforts to collect and disseminate information are not new. As a review by Anderson and Seltzer (2009) suggests, "the roots of the modern concept of federal statistical confidentiality can be traced directly back to the late nineteenth century" (p. 8). Notwithstanding this history, the literature on statistical disclosure methods is fairly recent by modern standards (Dalenius, 1977, is considered the seminal paper). In 1975, the U.S. Federal Committee on Statistical Methodology (FCSM) was organized by the Office of Management and Budget (OMB) to investigate issues of data quality affecting federal statistics. As part of this effort, the Subcommittee on Disclosure Limitation Methodology, created within the FCSM, published its 1994 Statistical Policy Working Paper 22 (SPWP22). This paper, which was revised in 2005 by the Confidentiality and Data Access Committee (CDAC, 2005), sets good practice guidelines and recommendations for all agencies regarding confidentiality protection.

Defining Confidentiality and Disclosure

A definition of confidentiality is given in SPWP22. According to the document, the definition endorsed by the President's Commission on Federal Statistics states that "[Confidential should mean that the dissemination] of data in a manner that would allow public identification of the respondent or would in any way be harmful to him is prohibited, and that the data are immune from legal process." This definition originates from the book *Private Lives and Public Policies* by Duncan, Jabine, and de Wolf (1993). Similarly, and according to the same source, "confidentiality differs from privacy" because "it applies to business as well as individuals. Privacy is an individual right whereas confidentiality often applies to data on organizations and firms."

Several different definitions of disclosure risk have been proposed. SPWP22 follows Duncan et al. (1993): "Disclosure relates to inappropriate attribution of information to a data subject, whether an individual or an organization." The same authors distinguish three types of disclosure: (a) when a data subject is identified from a released file (identity disclosure), (b) when sensitive information about a data subject is revealed through the released file (attribute disclosure), or (c) when the released data make it possible to determine the value of some characteristic of an individual more accurately than otherwise would have been possible (inferential disclosure).

Need for Protection, Need for Information

The biggest policy tension underlying the debate in the literature is the need to balance two inherently competing goals: need for information and need for protection. Federal agencies are required by law to protect the confidentiality of individual information. For instance, Title 13 of the U.S. Code prevents the census from releasing data in which any particular individual or establishment can be identified. Other legislation aimed at preventing such disclosures includes the Health Insurance Portability and Accountability Act (HIPAA) and the Confidential Information Protection and Statistical Efficiency Act (CIPSEA). At the same time, the government is committed to providing data to the research community for the advancement of knowledge (e.g., Open Government Directive).

On one hand, access to microdata (i.e., data collected on individuals or households) generates "more and better research, higher transparency, better assessment of data quality, better assessment of data gaps, and replicability" (Lane, 2007). Advocates of greater access to health care data justify their position based on the public's need for information about "quality, efficiency, effectiveness, and safety of the health care they receive and pay for" (Rosenbaum, 2010). On the other hand, scholars highlight the need to "protect patient privacy, the confidential nature of the patient/professional relationship, and health information security" (Rosenbaum, 2010). Data breaches also represent legal and financial liabilities for data custodians and erode public trust in their ability to handle data (Couper, Singer, Conrad, & Groves, 2008; Couper, Singer, Conrad, & Groves, 2010).

The disclosure literature is divided into two competing research paradigms: (a) that it is possible to protect and release data and (b) that privacy and confidentiality cannot be achieved in an environment in which personal information is gathered at an increasing rate by multiple people with multiple interests. The literature on disclosure limitation techniques and their achievements is extensive. In 1998, for example, the *Journal of Official Statistics* devoted an entire issue to this question. In addition, every 2 years the UNESCO sponsors an international conference (i.e., Privacy in Statistical Databases) that gathers worldwide experts from different disciplines to discuss current issues in the field. Proceedings are published by Springer in the series Lecture Notes in Computer Science (Domingo-Ferrer,

2002; Domingo-Ferrer & Franconi, 2006; Domingo-Ferrer & Magkos, 2010; Domingo-Ferrer & Saygin, 2008; Domingo-Ferrer & Torra, 2004; European Communities, 1998).

More recently, the book by Duncan, Elliot, and Salazar-González (2011) provides a comprehensive understanding of the principles and practice of statistical confidentiality. Other recent technical reviews are those by B. Chen, Kifer, LeFevre, and Machanavajjhala (2009) and Fung, Wang, Chen, and Yu (2010) on the latest developments in the field of Privacy-preserving data publishing, and by Fayyoumi and Oommen (2010) on statistical disclosure control and microaggregation techniques for secure statistical databases.

The belief behind the statistical literature is that it is, indeed, possible to minimize the risk of disclosure and, therefore, to release data to the public. This belief is, however, not shared by other scientists. For instance, computer scientists Narayanan and Shmatikov (2010) criticize such an approach by suggesting that the underlying assumption behind de-identification techniques is that personally identifiable information is a fixed set of attributes such as names and contact information. This, according to the same researchers, "creates the fallacious distinction between 'identifying' and 'non-identifying' attributes." In particular, these authors clarify that such a distinction might make sense in the context of one attack but is increasingly meaningless as the amount and variety of publicly available information about individuals grow exponentially (Narayanan & Shmatikov, 2010).

In a similar vein, computer scientists Cynthia Dwork and Moni Naor show that the type of privacy defined by Dalenius in 1977 ("access to a statistical database should not enable one to learn anything about an individual that could not be learned without access") cannot be achieved in general. The obstacle, according to these authors, "is in the auxiliary information." The main result in their paper is that "in any 'reasonable' setting there is a piece of information that is in itself innocent, yet in conjunction with even a modified (noisy) version of the data yields a privacy breach" (Dwork & Naor, 2010, p. 93). To sidestep this issue, Dwork (2006) defines *differential privacy* and shows that this type of privacy can be implemented and formally proven.

As pointed out by one of our reviewers, while both approaches (favored by statisticians or by computer scientists) balance utility and disclosure risk, differential privacy is limited in practice. It can only be used in settings where access to the data is remote and controlled. This removes traditional public use files (PUFs) or any microdata delivery from its scope. Consistent with our objective of providing a useful literature review for practitioners interested in making data sets publicly available, we concentrate our literature review on statistical disclosure methods.

Privacy Compromises

Privacy compromises in published data have been documented by Sweeny (1997), Agrawal and Srikant (2000), Algranati and Kadane (2004), and Ochoa et al. (2008). Recently, in a widely cited paper, Narayanan and Shmatikov (2008) show the feasibility of large-scale reidentification using movie-viewing histories. These authors suggest that statistical de-identification techniques provide only a weak form of privacy (Narayanan & Shmatikov, 2010) and that privacy protection has to be built and reasoned about on a case-by-case basis (Narayanan & Shmatikov, 2010).

Last, other scholars suggest that while attempts to reidentify individuals are partly mitigated through legal barriers, such as data user agreements or confidentiality agreements that explicitly ban users from doing it, "such policies provide no formal privacy protection guarantees" (Loukides, Denny, & Malin, 2010).

Consequences of Misapplication of Disclosure Avoidance Procedures

Although Winkler (2007), using artificial data, warned of "the severe analytic distortions of many widely used masking methods that have been in use for a number of years," not until recently have researchers documented problems with data released in PUFs.

In particular, Alexander, Davern, and Stevenson (2010) discover and document errors in public use microdata samples ("PUMS files") of the 2000 census, the 2003-2006 American Community Survey (ACS), and the 2004-2009 Current Population Survey, due to the misapplication of disclosure avoidance procedures. The particular procedure that caused the problem is not cited in the paper, nor disclosed in ACS's errata notes #47 or #50.

These authors show that for women and men ages 65 and older, age- and sex-specific population estimates generated from the PUMS files differ by as much as 15% from counts in published data tables. Additional analysis of labor force participation and marriage rates in the same publication shows that PUMS samples are not representative of the population at individual ages for those ages 65 and older, and that PUMS files substantially (a) underestimate labor force participation for those near retirement age and (b) overestimate it for those at older ages.

Finally, these authors emphasize that the problem could affect a whole range of stakeholders: researchers, social service agencies that rely on the data for policy research, and survey researchers who use PUMS data to generate population estimates.

De-identification Methods

The first step to prevent disclosure is to remove all direct identifying variables, such as name, phone number, and address. This step is intuitive, and one could naively think the data set is then safe from disclosure because no individual is explicitly identifiable. However, as cited earlier (Agrawal & Srikant, 2000; Sweeney, 1997) removal of direct identifiers does not protect all individuals from data disclosure or reidentification. A combination of just a few indirect identifying variables (such as birth date, gender, and zip code) can be used

to identify a large portion of individuals on any data set. These variables could then be matched to publicly available data to identify records. Disclosure risk represents the risk of indirect identifiers (IVs) being used to match records to an external data source that contains direct identifiers. The challenge of any de-identification technique is to limit reidentification via the use of indirect identifying variables.

This section describes the methods used to de-identify a data set. First, we define disclosure and disclosure risk and discuss the goals of disclosure treatment. We then describe the methods for de-identification of microdata and the use of nonsynthetic and synthetic treatments. Finally, we discuss the issues related to de-identification of longitudinal microdata.

Disclosure and Disclosure Risk Defined

Disclosure is the communication, either directly or by inference, of information about a member of a data set that could not be known without viewing the data set. We call this information "sensitive information" going forward. It is the obligation of the owner of the data set or data provider to prevent disclosure of sensitive information about members of the data set.

An intruder is someone who wants to discover sensitive information about any person (or other entity) from the data set. In short, disclosure is the process of an intruder discovering sensitive information about a target they did not know prior to intrusion.

Disclosure risk is a measure of the probability of disclosure for either an individual record or a data set as a whole. Skinner (2009) gives two useful definitions: (a) Disclosure risk is concerned with the possibility that the intruder will be able to determine a correct link between a microdata record and a known unit and (b) disclosure risk might be defined as the probability of disclosure with respect to specified sources of uncertainty, such as whether the disclosed information is accurately attributed to a target.

These definitions show that, for any data set, the calculation of disclosure risk relies heavily on assumptions about the intruder's knowledge. The more knowledge an intruder has about a potential target in a data set, the higher the probability that the intruder will be able to correctly identify the target and disclose information.

What Are the Goals of Disclosure Treatment?

Dalenius (1977) states that "access to a statistical database should not enable one to learn anything about an individual that could not be learned without access." While this is a noble goal, Dwork (2006) states that this level of privacy, zero disclosure risk, cannot be achieved for microdata or even published macrodata. Given that disclosure risk cannot be zero for analytically useful data sets, the goal should be to make the risk as low as possible. For instance, Skinner (2009) says that "confidentiality of the answers provided by a respondent might be said to be protected if the disclosure risk for this respondent and the respondent's answers is sufficiently low." Previous research (Dalenius, 1988; Fienberg & McIntyre, 2005) argues that data should be released if the probability of identifying an individual or entity in the data file is appropriately small. The emerging consensus of the field is that if the disclosure risk is small then data should be released; however, current research has not been specific about a defined risk measure, assumptions about the intruder, or what constitutes "small."

Winkler (1997) takes a different stance, stating, "in producing confidential public-use data files, statistical agencies should first assure that the files are analytically valid." That is not to say that data confidentiality is unimportant; it just calls out that producing a data set with low disclosure risk but little analytic utility is akin to not producing a data set at all because its release has no analytic benefit.

Therefore, the goal of the data producer should be to produce an analytically useful data file, with acceptably small disclosure risk for the individuals in the file. Currently, the definition of what is "acceptably small" is up to the provider, based on the provider's obligation to the participants in the data set. If risk cannot be made adequately small while preserving data utility, the provider should consider alternative methods of publishing or limiting access to the data. A combination of de-identification, access control, and data use agreements may be a more appropriate solution (Abowd & Lane, 2004; Lane & Schur, 2010).

The following sections are discussed in the context of treating a microdata file, but many of the same techniques can also be applied to tabular data (Skinner, 2009).

Nonsynthetic Treatments of Microdata

As opposed to synthetic disclosure treatments, where all records in a data set are treated, nonsynthetic methods treat only a fraction of the records in the data set. This is usually done deterministically to reduce the disclosure risk of a small group of (or single) records. Nonsynthetic disclosure treatments consist of three primary tools: global recoding, suppression, and perturbation. Nonsynthetic disclosure treatments are designed to specifically treat records with high disclosure risk to produce a data set that has a low risk of disclosure. However, the deterministic nature of the treatment can introduce selection bias that can degrade analytic utility (Singh, 2009).

Global Recoding and Local Suppression (GRLS)

Global recoding is a process of reducing the number of values a single variable can have in a data set. For example, an individual's birth date may exist in a data set and could be used as an indirect identifying variable. However, the

variable could be recoded to birth year and be less useful as an IV; it could be further recoded to 5-year intervals to make it even less identifying. Decisions on the appropriate level of recoding are based on the trade-off made between the competing needs for data confidentiality and preservation of data utility.

Local suppression is the process of removing, or suppressing, data from a data set. This can be done for single variables within a record or an entire record. Level of suppression is again determined by the de-identification strategy being used and the competing needs for data confidentiality versus data utility.

Sweeney (2002) defines k-anonymity as follows: "A release provides k-anonymity protection if the information for each person contained in the release cannot be distinguished from k-1 individuals whose information also appears in the release." The concept of k-anonymity is the same as "cell size," which was already in use by professional statisticians interested in limiting disclosure of information in public data sets. Willenborg and de Waal (1996) provide a historical perspective on this broadly used concept. Sweeney achieves k-anonymity in a data set through the use of global recoding. Indirect identifying variables are recoded until each combination of recoded variables has at least k number of records associated with it. At this point, no individual in the data set can be identified with certainty because no individual has a unique IV profile.

There are algorithms whose purpose is to make the process of recoding as efficient as possible by minimizing the amount of information loss while reducing disclosure risk. El Emam et al. (2009) discuss some of these information loss metrics in detail. Note that they are only useful in making decisions regarding recoding and suppression; they do not give the user/analyst any measure of data utility.

The concept of k-anonymity drives several real world systems, including Datafly, k-Similar (Sweeney, 2002); Samarati, Incognito, and Optimal Lattice Anonymization (OLA; El Emam et al, 2009); and µ-argus (Hundepool et al., 2008). Most of these packages use local suppression in addition to global recoding to create a k-anonymous data set.

Perturbation

Perturbation, another process that can be used to reduce disclosure risk, alters the values of variables on the data set. This could be performed to make reidentification more difficult on (a) variables expected to be known to the intruder (IVs) or (b) particularly sensitive information not known to the intruder (sensitive variables or SVs). Nonsynthetic perturbation treats only a portion of the records on the data set. Synthetic perturbation treats all records in the data set.

Nonsynthetic perturbation can be random or selective. Selective perturbation deterministically selects records for treatment to reduce disclosure risk. Also called blank and impute, this method selects values from single records, removes them from the record, then imputes a new value. This means that certain values may be targeted more frequently, creating a bias that is difficult for the analyst to quantify when interpreting the data (Skinner, 2009).

Data swapping was one of the first perturbation methods, proposed by Dalenius and Reiss as early as 1982. It was proposed as a method to transform a data set by exchanging values of SVs in such a way as to preserve their confidentiality while maintaining data utility. Records were selected for a "data swap" of a single SV if the swap resulted in a decrease in disclosure risk and preservation of marginal counts associated with that SV. This method was shown to reduce disclosure risk while protecting data utility for contingency tables and log-linear models; however, data utility was found not to be preserved for other types of analysis (Fienberg & McIntyre, 2005).

Substitution is similar to data swapping in that data from one record are replacing data on another record. As proposed by Singh (2009), substitution is the process of replacing some or all IVs in a record with the IVs from another record. It is different from data swapping in that the data only move in one direction. The pairing of records for this substitution process relies on techniques common in survey sampling for the imputation of missing values. Substitution reduces disclosure risk by creating uncertainty about the association between the IVs and SVs for a given record. Records can be selected for substitution deterministically to reduce risk, although in Singh, they are selected randomly.

GenMASSC: Global Recoding + Random Perturbation + Random Suppression + Calibration

Singh (2009) and Singh, Yu, and Dunteman (2003) propose the combination of multiple elements from synthetic and nonsynthetic frameworks to de-identify data while simultaneously controlling disclosure risk and information loss. The first treatment step is global recoding, the amount of which is driven by reducing disclosure risk within constraints that preserve data utility. The second step is random perturbation by substitution. Records are randomly selected to have their indirect identifying variables replaced with variables from a different record on the data set with similar properties. The third operation is the random suppression of entire records from the data set. The rates of selection for substitution and suppression can be functions of the disclosure risk of the record, so that records at higher risk may be chosen for treatment at a higher rate. After all treatment is complete, the data set is calibrated so that predetermined analytic values are representative of the data set prior to treatment. The stochastic nature of this treatment limits the amount of bias in the treated data set and allows the data provider to monitor and control the amount of bias and variance in the treated data for a given level of disclosure risk.

Synthetic Treatments of Microdata

Synthetic treatments treat all records in the data set to create a new, "synthetic" data set that is representative of the original data file. This is usually done by treating all the indirect identifying variables (such as birth date, gender, and zip code) for each record in the data set. The indirect identifying variables may be changed by a variety of methods, including perturbation, multiple imputation, and other model-based techniques (Skinner, 2009). Such methods are tuned to preserve data utility, at least for anticipated analyses, and data confidentiality is assumed to be improved because none of the records in the microdata represents an actual individual.

In stochastic perturbation, indirect identifying variables are modified by a random mechanism. Continuous variables are altered by adding random noise. The noise may be added so that the mean and variance of the variables within certain domains will be preserved. However, correlations outside the specifications (i.e., variables selected for synthetic treatment) at the time of the noise addition will not be preserved (Skinner, 2009). For instance, consider the case when a data producer applied synthetic treatment aimed at preserving the relationship between income and gender but decided to leave race outside the specification. If a user were to analyze the relationship between race and income, the data might be distorted because the treatment was not trying to preserve that particular relationship (i.e., race and income). In general, it may not be practical, or possible, for all relationships to be preserved during treatment. The data producer may have to make some tough decisions about which relationships are most important and, thus, to be preserved during treatment.

Fuller (1993) discusses methods that can be used to improve the analytic utility regarding these "unspecified" variables. These methods require the data provider to let the analyst know the standard deviation of the noise that was added to the indirect identifying variables. The analyst must then add procedures that use this information to the analysis for the output to gain the same inferences as the untreated data. A potential difficulty with this approach is that some reidentification experts suggest that knowing information about how the noise was applied to the variables can allow an intruder to reverse "engineer" the data file and potentially identify individuals. Details on how noise can be reverse engineered are available in Domingo-Ferrer, Sebe, and Castella-Roca (2004).

Other forms of synthetic perturbation include data swapping or substitution for all records in the data set. This is an extension of the previously mentioned nonsynthetic perturbation methods, where selection for treatment is expanded from deterministic or random selection processes to 100% selection as part of a synthetic treatment.

Categorical variables can be reclassified using a modeling mechanism such as the postrandomization method (PRAM; Gouweleeuw, Kooiman, Willenborg, & De Wolf, 1998). These methods use other variables in the data set to find probabilities for levels of the categorical indirect identifying variable that is to be treated. The model is then scored for all records and the initial value of the variable replaced with the new variable. Depending on the structure of the model used, the treated data set can be analyzed as is, or may require additional information from the data provider to the researcher to perform valid analysis. As with noise addition, the application of this extra information requires increased computations by the analyst to produce valid results.

Multiple imputation uses a model to create synthetic records based on a known distribution of indirect identifying variables for the data set. More records are created through these processes than are intended for release. The population of synthetic records is then sampled multiple times to estimate the analytic properties of the data set, with one of the samples released as the treated data set. The analyst can treat this data set as a survey sample with known variances and use standard survey sample techniques of analysis. A detailed review of this process can be seen in Rubin (1993). Abowd, Stinson, and Benedetto (2006) present an implementation of this technique using linked data from the census, Social Security Administration, Internal Revenue Service, and Congressional Budget Office. Multiple imputation removes the need for the provider to pass information regarding perturbation to the analyst, which also frees the analyst from extra calculations required when analyzing data.

Synthetic data approaches can also be applied to a subset of the data (Little & Liu, 2003; Reiter, 2009). For instance, an agency could be interested in replacing income when it exceeds a certain threshold but is willing to release all other values (Reiter, 2009). The result is a partially synthetic data set.

As mentioned, the advantage of synthetic data is that they are designed to preserve data utility. The data confidentiality of the process is assumed to be implied because no "real" records are released. Domingo-Ferrer & Torra (2003) shows that this is not necessarily the case, however, and that there are reidentification techniques capable of disclosing information about individuals in a synthetic data set that has been de-identified using synthetic treatments. Other limitations of synthetic microdata are (a) the expertise and effort required to build a model and (b) that the quality of the treated data and its analysis is a direct result of the quality of the model (Singh, 2009; Winkler, 2007).

De-identification of Longitudinal Microdata

The de-identification of longitudinal data has not been explored by many researchers, as it has been thought an unobtainable goal. Several researchers note that preserving the data confidentiality and data utility of a public use longitudinal data set may be inherently incompatible goals (Abowd & Woodcock, 2002; Nadeau, Gagnon, & Latouche, 1999). However, Abowd and colleagues have successfully used multiple imputation synthetic de-identification techniques to treat longitudinal data (Abowd et al., 2006; Abowd & Woodcock, 2002). It is

important to note that these techniques summarize the longitudinal data prior to treatment and data are still de-identified by individual, not by longitudinal event or record. In addition, data present on the longitudinal file that were not summarized prior to treatment are not available to the analyst, as there is no direct publication of the longitudinal data.

Assessing Data Utility After Treatment

Data anonymity and analytic utility are in constant tension, with increases in one resulting in a decrease in the other. However, assessment of data utility is a vital step following de-identification treatment and should go in concert with disclosure risk analysis. Researchers have developed a variety of methods by which to automate the analysis of data utility using specially designed software. Together with the assessment of risk, it has also been shown that certain methods of de-identification are less effective than others at maintaining data utility and protecting personal privacy (Winkler, 2007). Kennickell and Lane (2007) give a good overview of the role of data utility in the context of de-identification treatment and several methods that appear throughout the literature.

One such method of comparison is between simulated research results on treated and untreated data sets. In their analysis of disclosure risk and analytic utility, Brickell and Shmatikov (2008) show it is necessary to render a data mining utility near useless to researchers when using generalization and suppression of quasi-identifiers to de-identify a data set. These researchers compared their results with what was called "trivial sanitization" of the data set, which simply omits either all quasi-identifiers or all sensitive attributes to provide maximum privacy.

Rastogi, Suciu, and Hong (2007) depict a framework for describing privacy and utility of a de-identified data set. Privacy is defined as a comparison of an attacker's probability of an ordered list of elements against the probability based on experimental observation. To illustrate data privacy and utility, these researchers use census data to evaluate a selection of queries with up to three attributes and estimates of the error bound on counting queries. They describe a simple anonymization algorithm that uses random insertions and deletions of varying series of data in, or from, the database.

Other methods compare estimates taken from treated and untreated data sets. Winkler (2007); Raghunathan, Reiter, and Rubin (2003); and Abowd et al. (2006) all use comparisons of correlation and regression coefficients to assess the analytic utility of data sets treated to reduce disclosure risk. These methods assume the data provider can anticipate many of the correlations that will be useful to the user and measure the impact of treatment on these relationships prior to release.

Singh (2009) assesses data utility of a treated data set by comparing the means of several variables of interest across multiple replications of treatment. By using multiple treated data sets, a relative (to the mean of the untreated data) root mean squared error (RRMSE) can be computed to describe how much the value in a treated data set can be expected to vary from the value in the untreated data. This measure of data utility is on the same scale for all variables of interest because it represents the error of the treated values relative to the untreated values. The maximum RRMSE represents the error of the least reliable variable, or relation of interest, and can be used as a simple measure of data utility. This method also assumes that the data provider has a good idea of the relationships that will be important to the user to measure data utility in the proper context.

Reidentification Methods

With the collection and provision of data comes the risk of identifying individuals within data sets and the associated harms that can run the gamut from inconsequential to catastrophic. Several methods have been developed to assess the risk of this reidentification and test data sets for the ability to identify specific people. The literature reveals four areas by which reidentification practice occurs: linking records across multiple data sets, linking data across multiple data sets, matching patterns within multiple data sets, and, most recently, identifying individuals in the public space from usage patterns (Winkler, 2004a, 2004b).

The methods presented in the following sections all seek to identify individuals when one or more data availability scenarios are present (Domingo-Ferrer & Torra, 2003):

- Where there are common variables and a common terminology in multiple data sets and these are leveraged to effect reidentification of individuals;
- Where there may be common variables but differing terminology between data sets;
- Where there are no variables in common between comparable data sets but an existing and common terminology exists; and
- Where, in the final and most challenging scenario, variables and terminology are different between data sets.

In each, the method used makes at least the assumption that there are individuals in common within the associated data sets.

Assessing Risk

Generally, disclosure risk for a target increases as more is known, in terms of quantity and precision of data. One of the most common methods to measure disclosure risk is to count the number of unique records within a data set with a limited set of individual record characteristics (El Emam et al., 2010). Research has also focused on estimating the number of uniques within a population from a sample of data given different possible population distributions (Bethlehem, Keller, & Pannekoek, 1990; G. Chen & Keller-McNulty, 1998).

The challenge in developing different methodologies to assess risk is the need to accurately reflect risk. Too conservative risk assessment needlessly sacrifices data utility in favor of individual anonymity; risk assessment that errs on the other side risks disclosure of sensitive information. This has been described as the over- and under fitting of risk estimates and is the main thrust of research in risk disclosure and development of the two-way interaction model (Skinner, 2007).

Skinner and Shlomo (2008) further refine the two-way interaction model for estimating disclosure risk measures through development and use of diagnostic criteria for model choice with the goal of balancing over- and under fitting. These researchers illustrate the ability to use Poisson log-linear models in the assessment of risk in large and sparse contingency tables spanned by key variables. Their approach has been shown to be useful for file-level and record-level measures of risk.

Truta, Fotouhi, and Barth-Jones (2004) introduce a general framework for assessing disclosure risk by classifying data set attributes based on either potential identification utility or order with regard to domain of value. These values, termed change factors, measure the magnitude of masking applied to data and the modification that has occurred to key attributes. Using simulated medical billing data with identifier attributes removed, the researchers are able to show minimum, maximum, and weighted disclosure risk values for a number of different masking methods. They perform a series of experiments whereby random noise is added to identifying attributes (age, sex, ZIP code, and billed amount) and the effect on disclosure risk measured. The method described by these researchers allows for a measure to assess the amount of information loss as a result of the specific masking method used; it also presents a way to measure and set the level of masking desired to achieve a preset level of risk.

Benitez and Malin (2010) illustrate the wide gap between perceived threats of reidentification and actual results. The paper tests voter registration data as a route of potential reidentification of publicly released health records protected by the Safe Harbor policy. In particular, the authors suggest that allusion to the potential uses of voter lists in the literature (Sweeney, 1997) rarely acknowledges the complexity of the data (i.e., access and quality) or the economic costs to an attacker.

Their risk analysis estimation in Benitez and Malin (2010) is probabilistic in nature, as it quantifies the likelihood of reidentification for each member of a group. The analysis consists of a three-step process: (a) determine the fields available to an attacker (i.e., year of birth, race, and gender in health records and date of birth, year of birth, race, gender, and county of residence in voter registration); (b) group census data according to these fields to estimate population counts; and (c) add results obtained by applying risk estimation metrics to the results, and normalize by total population.

Benitez and Malin (2010) find that risk levels and costs vary widely across different states due to individual voter registration policies, for example, with more permissive states having higher risks of disclosure.

Individual risk methodology, developed by Benedetti and Franconi (1998), involves the computation of individual risk for each unit of analysis within a data set as the probability of correct reidentification. The risk of reidentification is expressed through the concept of unique or rare combinations in the data, and the methodology uses sampling weights to account for the uncertainty of whether such unique combinations are common or rare in the population. All records with individual risk above a fixed threshold are defined as being at risk, implying that disclosure control methods must be used to protect these records.

Elliot (2000) defines an additional measure of disclosure risk that measures correct matches between actual and masked data sets. Termed Data Intrusion Simulation (DIS), the researcher describes a method that uses the target data set to estimate matches given a unique match. The method forgoes the use of an entire population and instead uses a sample. The method contains five steps:

1. Take a sample microdata file with sampling fraction.
2. Remove a small random number of records to make a new file.
3. Copy back a random number of the records.
4. Match a simulated fragment of the identification file with the target microdata file. Generate the probability of a correct match given a unique match for the fragment.
5. Iterate until the estimate stabilizes.

The effect is to generate a risk of disclosure for a given data set without assuming that a given unique record is a population unique. In addition, this method retains the usefulness of matching against actual data without being a nongeneralizable, ad hoc approach using the entire data set on which to attempt a match.

Sources of Data for Reidentification

There are many entities, not covered by HIPAA, that collect and disseminate identifying information to clients and other users. This information is collected from a variety of sources and, if used in combination with information from health data sets, may potentially contribute to the identification of individuals. To our knowledge, there have not been demonstrated reidentification attacks using these sources. Such sources of data and information include social networking websites, transactional data, voter registration records, state agencies, and web crawlers, among others.

- *Social networking sites* collect a plethora of identifying information including data on the habits and behaviors of consumers, for instance, websites such as patientslikeme.com, healthboards.com, and

WebMD.com. There are many examples of this information already being used in a commercial manner—for example, in targeted online advertisement. However, it is worth noting that this is not evidence that the data are used for reidentification purposes.

- *Transaction data*, such as collected by credit card companies or credit-reporting companies (e.g., Equifax, TransUnion, and Experian), hold enormous amounts of sensitive financial transactional data that could potentially be recombined with publicly released data to reidentify consumers and sold for behavior prediction and targeted marketing.
- *Public information*, such as voter registration, court cases, and many other government transactions, is publicly available and aggregated by private companies. Examples include Intelius, NextMark, and Infogroup (previously InfoUSA).
- *Health care data* are also becoming increasingly available. For instance, states such as Vermont and Texas have de-identified administrative data on hospital discharges available either free (Vermont) or for a fee (Texas). Similarly, the Pennsylvania Health Care Cost Containment Council provides identified tabular data on its website for free at the provider level (e.g., hospital, medical doctor) and microdata for a fee.
- *Internet search engines*, such as Pipl.com, are also proliferating. Unlike most data aggregators, these sites crawl the web looking for other websites and data miners with personal data. Interested users only need to provide the search engine with the most they know about the person they are looking for (e.g., first name, last name, city, state). In response, the search engine displays links to information available on the web for persons with matching characteristics.

In what follows, we present selected methods that have been developed either as modifications of methods originally intended for other purposes or specifically for data reidentification. Reidentification methods expose weaknesses in masking methodologies and other efforts to protect individual and group privacy. We do not present an exhaustive list of the many available reidentification techniques, which vary widely in their complexity. Rather, we provide a broad overview of the major operating themes they represent.

Record Linkage

Initially developed as a method to synchronize files in cases where one may contain incorrect or inaccurate data, record linkage seeks to use two or more lists to classify pairs and form definite matches between each to string together records from different data sets (Malin, Sweeney, & Newton, 2003). For record linkage to proceed, a number of assumptions must be made about data within the sets in question. One such assumption is that there are common variables between the files. Matching data sets against commercially and publicly available data is one method by which reidentification can occur (Winkler, 2004a). Increasingly sophisticated reidentification methods combined with greater availability of public information has resulted in increased risk of data disclosure (Winkler, 2004a).

Winkler (2004b) describes record linkage methods as using metrics to scale the ranges of variables while partially accounting for dependencies between them. Scheuren and Winkler (1997) have illustrated how economic variables can substantially increase the accuracy by which linkages in administrative lists can be made. Correlations between these variables allow researchers to create predictors that permit records from one of the files to be closer to smaller subsets of other records in the other file. The probability of identifying individuals increases as the subset of predicted records decreases.

The most commonly used example in the literature is voter registration records, but many other data sets can be used. Loukides et al. (2010) illustrate how genetic research data can be used to reidentify patients within a health data set, even after suppression methods including the application of generalization and the HIPAA Privacy Rule. The authors achieve this by linking diagnosis codes (International Classification of Diseases–Ninth Revision [ICD-9]) derived from electronic medical records with released research data in the form of DNA sequences.

Bacher, Brand, and Bender (2002) illustrate the potential to reidentify persons within data sets using a feature of commonly used statistical software. Specifically, these researchers use cluster analysis with SPSS to match survey data against register data in a German context. The approach chosen by the researchers transforms and weighs variables and obtains a reidentification risk of approximately 10%.

Data Aggregation

While record and data linkage require direct relationships between features associated with the data sets, aggregation-related approaches attempt to reidentify when there are no common attributes (Winkler, 2004b). The objective of data aggregation with regard to reidentification is to create an ordering of the data using combinations of individual attributes. To do this in data sets containing numerical data, several assumptions are necessary, including the following:

1. There are common individuals in the two data sets.
2. The structures to the data contained within the data sets are similar.

Reidentification is then achieved by matching records that have similar groups of attribute combinations; it occurs when public information is linked to data files, and names,

addresses, or other information are at risk of being released (Domingo-Ferrer & Torra, 2003). When it is known that the populations in the acquired data sets are overlapping, it becomes possible to use variables from one of them to identify a subset of records from the other (Reiter, 2003).

Probabilistic Inference

With the use of Markov random fields and graph partitioning algorithms, the ability to increase the chances of identifying individuals through the linking of records and data in groups of files has been illustrated by McCallum and Wellner (2003). There are a number of important differences between data linkage and record linkage, particularly in that data linkage was developed with the intent of reidentification. The aim of data linkage is to make reidentification possible for data completely lacking seemingly identifiable information (Malin et al., 2003). Furthermore, attributes of the associated data sets are not required to be the same, as the technique makes use of inferential relationships between file attributes, which is the process of attempting to reidentify when there are no common attributes between data sources.

Narayanan and Shmatikov (2008) develop a general class of algorithms to identify individuals within large, sparse data sets (i.e., data sets where only a fraction of the cells contain relevant information). These algorithms take into account some amount of auxiliary information provided on a target, and score the records within the data set according to how well it matches the target. From here, a matching criterion is applied, and a single record or set of probable records is identified as a match. Narayanan and Shmatikov also illustrate the algorithms' resistance to de-identification data perturbation and methods of generalization and suppression. The algorithms described were applied to the Netflix Prize set up by the movie rental company to improve their system. More than 100 million customer movie ratings were made publicly available. Despite the removal of identifying customer information, the researchers were able to illustrate that simply removing identifying information is insufficient to produce anonymity.

In data linkage, characteristics of individual records of the data set are combined to estimate the uniqueness within a known population (Sweeney, 2000). Sweeney illustrated that, based on gender, ZIP code, and full date of birth, 87% of the U.S. population can be uniquely identified. The addition of extra information (i.e., race) adds more granularity and scarcity, thus increasing the likelihood that a record is unique. Sweeney has indicated that linkage is established through known attributes, and the probability of identifying individuals increases with the addition of further data (Sweeney, 2000).

It is important to note that these attacks were not on health data. The question then becomes whether this kind of inference can be applied to health data. We did not find any evidence of this in the published literature.

Trail Reidentification

Trail reidentification expands on the concept of reidentification by seeking to identify people who visited named locations in a network environment (Malin et al., 2003). Trail reidentification seeks to reconstruct a person through separately collecting and subsequently relating de-identified data on people who visited the location. The collected de-identified data consist of very few data fields. Recognizing uniquely occurring visit patterns across the de-identified and identified data sets provides the basis for trail reidentification. These observations are made explicit by constructing a matrix of shared de-identified data and a matrix of shared identified data. The relationship to health information data sets exists in the ability to use this trail reidentification information, separately or in combination, to locate individuals and associated sensitive health information (Malin et al., 2003). Information gained by way of trail reidentification may be leveraged with health data set information to further uncover health or chronic disease conditions.

Standards for Acceptance of a File as Safe in Health Care Data Sets

There are two elements to the HIPAA Safe Harbor method of de-identification: (a) 18 specific identifiers and (b) actual knowledge. The Safe Harbor method has two parts. Part I dictates the removal or coarsening of 18 direct, or almost direct, identifiers that may be present in any data set. These identifiers fall into four categories: names, dates, contact information, and record IDs. Part II of the Safe Harbor method requires the covered entity to ensure that it possesses no actual knowledge of an individual being at risk of disclosure after removal of the 18 identifiers.

The critical part of the aforementioned standard is its incorporation of a reasonable person standard. While not a defined legal term of art, this language likely indicates a significant safeguard for those who de-identify data. From the perspective of legal interpretation, language like "to which there is no reasonable basis to believe" indicates that so long as the covered entity was not negligent in the de-identification process, it is likely exempt from liability as long as it acts reasonably and does not believe that reidentification could occur.

In 2002, researchers at the Carnegie Mellon's Data Privacy Lab suggested the concept of "The Minimal Risk Standard" as a way to operationalize Part II of the method for commercial purposes (Sweeney, 2010b). Two companies, Privacert Gold Standard and Quintiles, licensed the Data Privacy Lab's risk-assessment technology to provide HIPAA certifications for de-identified data (Sweeney, 2010a, 2010b). According to the Minimal Risk Standard, the identifiability of proposed data should be no more than the identifiability if the proposed data adhered to Safe Harbor Part I (Sweeney, 2010b). This, in practice, became a question of measuring the risk of

reidentification of a data set under Safe Harbor Part I. Work by Sweeney (2000) based on demographic uniqueness in the U.S. population showed that Safe Harbor Part I provides a risk of reidentification of 0.04% when demographic information released is restricted to gender, year of birth, and county of residence.

The online appendix of El Emam (2011) includes a summary of metrics that have been used for identity disclosure in actual de-identification projects, including approaches embedded in software such as µ-argus.

Access

The vast amount of data now collected on human beings and organizations as a result of cyberinfrastructure advances has created significant opportunities for social scientists to study and understand human behavior. At the same time, technologies have recently emerged, such as virtual private network (VPN), biometrics, and virtual computing, that permit microdata to be accessed in convenient ways while also protecting data confidentiality (Lane, Heus, & Mulcahy, 2008). The legal framework surrounding data access has evolved in recent years on a parallel course. For instance, landmark legislation, the CIPSEA of 2002, establishes rigorous confidentiality safeguards while setting provisions for the statistical agencies to "designate agents, by contract or by entering into a special agreement" for the purpose of performing "exclusively statistical activities, subject to the limitations and penalties" within the boundaries stipulated in the confidentiality safeguards. As Bradburn and Straf (2003) argue, such laws foster norms that facilitate access to meaningful statistical records and protect respondent confidentiality.

Access Modality Options

Given these recent changes in technology and legal guidance, data producers have several data dissemination options from which to choose. These options vary considerably in disclosure risk, analytical utility of the data, and ease of access. And the different data access modalities may be used independently or in combination, depending on one's dissemination objectives and intended audience.

For example, data producers may release microdata via PUFs that provide access to anonymized versions of data sets. PUFs are widely accessible through CD-ROMs or the Internet, and given their broad reach, statistical agencies use techniques like variable suppression, top and bottom coding, noise infusion, and geographic aggregation before releasing PUFs to protect the confidentiality of the respondents (Weinberg, Abowd, Steel, Zayatz, & Rowland, 2007). Although such techniques are understandably required, they often diminish the usefulness of the microdata (United Nations, 2007) and, thus, are not the optimal dissemination tools in terms of analytic utility. Statistical agencies also release synthetic microdata. Importantly, however, all the benefit of synthetic data depends on the validity of the models used to create them.

Similar to PUFs, online tabulation engines and statistical data cubes provide another alternative to giving researchers full access to raw microdata. At the request of the researchers, most often online, such tabulation engines generate customized summary tables and matrices after having gone through an automated disclosure review process. Online tabulation engines are easily accessible through the Internet and retain confidentiality through suppressed summary tables; however, they are arguably less useful than PUFs for analytical purposes.

Remote batch processing is another dissemination modality. Although researchers do not have full access to the data sets, they submit programs or codes remotely via the Internet and receive their output once it has been reviewed for disclosure control by the statistical agency. The execution is generally done offline; thus, the process is not interactive. While most batch processing systems use filters to suppress certain queries and results, in the same way as PUFs, the output obtained from this access modality is still potentially more useful than that obtained through PUF data sets (Weinberg et al., 2007). A review of such remote batch processing arrangements shows, however, that while they are relatively secure and can be effective for smaller requests, they can be slow when large computation is required (United Nations, 2007). Also, the noninteractive aspect of this access modality can be a hindering experience for researchers.

The general theme that emerges from the aforementioned dissemination modalities is that there are serious trade-offs that need to be examined in terms of data access solutions, including data utility, confidentiality, security, and ease of use. While PUFs, remote batch processing, and tabulation engines are easy to access and have incorporated security measures to protect confidentiality, those measures limit analytic utility. There are, however, other options available to data providers that allow researchers to increase the analytical utility of the data.

Licensing is one such example. Under this method, approved researchers are granted a license via a contract to analyze restricted-use microdata (Weinberg et al., 2007), and access is provided through various means, such as CD-ROMs or remote access (United Nations, 2007). The U.S. National Center for Educational Statistics (NCES), for example, uses this method for a large number of its confidential data sets; so does the U.S. Bureau of Labor Statistics (BLS) for access to its National Longitudinal Surveys of Youth (Weinberg et al., 2007).

Due in part to recent congressional legislation and OMB guidance on data sharing, as well as increasing concerns that licensing alone cannot adequately protect data confidentiality, since 2006 U.S. Government agencies have explored new ways of disseminating information. For example, rather than simply "pushing out" microdata through licensing contracts

to researchers, the National Institute of Standards and Technology, U.S. Department of Agriculture, the National Science Foundation (NSF), and other data producers are currently "pulling in" researchers via secure remote access nodes to sensitive data housed in the NORC Data Enclave.

One problem is that licensing allowing researchers direct access to confidential microdata involves hundreds if not thousands of CDs with disclosive microdata being shipped across the United States; hence, each access node (i.e., contracted researcher), arguably, is a potential confidentiality concern. Although licenses legally bind researchers to maintain confidentiality, even a single breach can be damaging to the reputation and mission of the data producer. This concern is exacerbated by the fact that data in this model are delivered through mediums such as CD-ROMs that can easily change hands (challenging the chain of command), even without malicious behavior on the part of researchers. Therefore, for very sensitive microdata containing detailed personal identifiers, allowing this mode of access could be potentially risky, however trustworthy the individual researchers may be.

By sharp contrast, remote and physical data enclaves (also known as Research Data Centers or RDCs) are secure dissemination mechanisms. Whereas remote access platforms provide convenient access via an encrypted terminal session, RDCs only allow on-site access. To protect confidentiality, remote and physical data enclaves maintain stringent physical and computer security guidelines, preventing any results from being exported from the controlled environment without going through a formal disclosure review process.

An obvious advantage of remote and physical data enclaves is that researchers often have access to the most detailed version of the data, that is, raw microdata, devoid of suppression. Access to such analytically useful data through RDCs, however, comes at a price: They are very expensive to operate and are not convenient to all potential researchers (i.e., they require researchers to be physically present at the facility). Furthermore, the process for reviewing proposals or what results may be publicly released out of an RDC is very cumbersome and time-consuming (United Nations, 2007).

Remote access data enclaves typically use secure technologies, such as virtual computing, to allow approved researchers to connect to a data server that hosts the actual microdata and work in a remote-desktop environment. While users work in a familiar desktop environment, no output may leave the secured environment without first undergoing stringent statistical disclosure control. As pointed out by one of our reviewers, a determined intruder can find ways to overcome almost any obstacle. For instance, she can easily print a database to screen from the remote server and then capture the data using optical character recognition (OCR) technology. This is cumbersome and prone to error, but largely automatic and outside the control of the enclave administrator. Even if screen capture technology is not available, an off-site user can simply record the remote desktop with a camera. This is called the "analogue hole" in cryptography. Whether it is possible to apply stringent statistical disclosure to every single output, taking into account previously released outputs and future outputs is critical in this approach. This has not been proven in the literature.

Conclusion

Achieving data and information dissemination without harming any individual or any group is a central task of any entity in charge of collecting data. The balance lies in protecting the privacy of those in the data while minimizing data utility loss (Kinney et al., 2009). Although the need for such balance is true of every data set, it becomes more critical when the information collected is about personal health status and personal health care received.

Although several scientific disciplines have different views about the degree to which such balance can be achieved, they all agree that creating a file that is totally reidentification risk free is an impossible task. The question is, then, how much risk is tolerable. Statisticians offer several methodological approaches to balance disclosure treatment and utility, while accepting some level of risk. Some computer scientists are skeptical, and show low tolerance to risk, arguing for limiting the release of data to the public. Critics of microdata releases also point out that confidentiality agreements and data use agreements provide no formal privacy protection guarantees.

Research showing analytic distortions of widely used masking methods raises concerns about the misapplication of disclosure avoidance procedures. Besides embarrassment to the agency, the problem could affect a whole range of stakeholders.

Scientists have developed a variety of definitions and frameworks to quantify disclosure risk and a variety of methods to limit disclosure risk. These methods range from the simple suppression of a field or a subset of values in a field, to intricate perturbation methods such as data swapping and imputation via synthetic methods. Although some methods have been shown to be better at masking specific fields in a data set or to provide better protection while minimizing utility loss, the literature does not emphasize one method over another. Similarly, although different software applications have been designed and are available, there is no discussion in the literature about which one is best.

Our literature review shows that scholars have devoted considerable attention to the development of methods to mask microdata in settings in which units are not followed over time. However, the literature is sparse on longitudinal data. The addition of time as a variable adds a level of complexity that is still an open question in the field. The same conclusion holds for data utility metrics. Although several metrics have been proposed, the literature is vague regarding which one to use in practice.

Similarly, scientists have developed several methods to assess whether disclosure techniques have achieved the desired

protection. These include (a) record linkage, in which unique combinations of variables (e.g., gender, age, zip code) are used to match records in two or more data sets (e.g., a medical record and a voter registration list) and (b) probabilistic inference, in which sophisticated algorithms, taking into account some amount of auxiliary information provided publicly on a target, are able to detect with high probability a record or set of records of any given individual (i.e., identify whether a person with certain publicly known information is in a data set, and learn additional information about that person from the data set).

Such methods operate under different assumptions and have been validated empirically in very specific contexts. A common theme in the literature is the threat imposed by the growing amount of auxiliary information available either free or at very low prices, and the possibility that the monetary cost of an attack decreases with the availability of data and the growth in computer power. In addition, despite efforts to establish standards for acceptance of a data set as safe for public release, methods for quantifying the risk of reidentification are scarce in the literature.

Although there is abundant theoretical and empirical research, our review reveals lack of consensus on fundamental questions for empirical practice: how to assess disclosure risk, how to choose among disclosure methods, how to assess reidentification risk, and how to measure utility loss. As stated in Kinney et al. (2009), "it is not known whether the choice of measures is a problem with theoretical or methodological structure or merely disconnected special cases amenable only to empirical analysis" (p. 132).

Access to microdata has also received attention in the literature. Modalities vary in terms of disclosure risk, analytical utility, and ease of access. Several authors make interesting cases about the advantages and disadvantages of each method. But more empirical research is needed.

Acknowledgments

We thank Chris Haffer from the CMS Center for Strategic Planning, and Craig Coelen, Avi Singh, Daniel Barth-Jones, and two anonymous referees for their thorough review of and comments on this article. The views expressed in this article are those of the authors and do not necessarily reflect the views of the U.S. Department of Health and Human Services, the Centers for Medicare and Medicaid Services, IMPAQ International, NORC at the University of Chicago, or Optimal Solutions Group.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research and/or authorship of this article: The research in this article was supported by the Centers for Medicare and Medicaid Services under contract number 500-2006-00007I/#T0004 with IMPAQ International.

References

Abowd, J. M., & Lane, J. (2004). *New approaches to confidentiality protection: Synthetic data, remote access and research data centers*. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.85.3083

Abowd, J. M., Stinson, M., & Benedetto, G. (2006). *Final report to the social security administration on the SIPP/SSA/IRS public use file project*. Retrieved from http://www.bls.census.gov/sipp/synth_data.html

Abowd, J. M., & Woodcock, S. D. (2002). Disclosure limitation in longitudinal linked data. In P. Doyle, J. I. Lane, J. J. Theeuwes, & L. V. Zayatz (Eds.), *Confidentiality, disclosure, and data access* (pp. 215-278). Amsterdam, Netherlands: North Holland.

Agrawal, R., & Srikant, R. (2000). Privacy-preserving data mining. In *Proceedings ACM SIGMOD International Conference on Management of Data* (pp. 439-450). Dallas, TX.

Alexander, T., Davern, M., & Stevenson, B. (2010). Inaccurate age and sex data in the census PUMS files: Evidence and implications. *Public Opinion Quarterly*, *74*, 551-569.

Algranati, D., & Kadane, J. (2004). Extracting confidential information from public documents: The 2000 Department of Justice Report on the federal use of the death penalty in the United States. *Journal of Official Statistics*, *20*, 97-113.

Anderson, M., & Seltzer, W. (2009). Federal statistical confidentiality and business data: Twentieth century challenges and continuing issues. *Journal of Privacy and Confidentiality*, *1*, 7-52.

Bacher, J., Brand, R., & Bender, S. (2002). Re-identifying register data by survey data using cluster analysis: An empirical study. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, *10*, 589-607.

Benedetti, R., & Franconi, L. (1998). Statistical and technological solutions for controlled data dissemination. In *Pre-proceedings of New Techniques and Technologies for Statistics*, *1*, 225-232. Retrieved from http://neon.vb.cbs.nl/casc/Software/MuManual4.2.pdf

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

Bethlehem, J. G., Keller, W. J., & Pannekoek, J. (1990). Disclosure control of microdata. *Journal of the American Statistical Association*, *85*, 38-45.

Bradburn, N., & Straf, M. (2003). The eleventh Morris Hansen lecture information and statistical data: A distinction with a difference. *Journal of Official Statistics*, *19*, 321-331.

Brickell, J., & Shmatikov, V. (2008). The cost of privacy: Destruction of data-mining utility in anonymized data publishing. In *Proceedings of 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)* (pp.70-78). Retrieved from http://www.cs.utexas.edu/~shmat/

Chen, B., Kifer, D., LeFevre, K., & Machanavajjhala, A. (2009). Privacy-preserving data publishing. *Foundations and Trends in Databases*, *2*, 1-167.

Chen, G., & Keller-McNulty, S. (1998). Estimation of deidentification disclosure risk in microdata. *Journal of Official Statistics*, *14*, 79-95.

European Communities. (1998). Statistical data protection '98. In *Proceeding by Office of Official Publications, Statistical Office of the European Communities*. Retrieved from http://epp.eurostat.ec.europa.eu/portal/page/portal/eurostat/home

Confidentiality and Data Access Committee. (2005). *Federal committee on statistical methodology* (Statistical Policy Working Paper No. 22). Retrieved from http://www.fcsm.gov/committees/cdac/cdac.html

Couper, M., Singer, E., Conrad, F., & Groves, R. (2008). Risk of disclosure, perceptions of risk, and concerns about privacy and confidentiality as factors in survey participation. *Journal of Official Statistics*, *24*, 255-275.

Couper, M., Singer, E., Conrad, F., & Groves, R. (2010). Experimental studies of disclosure risk, disclosure harm, topic sensitivity, and survey participation. *Journal of Official Statistics*, *26*, 287-300.

Dalenius, T. (1977). Toward a methodology for statistical disclosure control. *Statistik Tidskrift*, *15*, 429-444.

Dalenius, T. (1988). *Controlling invasion of privacy in surveys*. Stockholm: Statistics Sweden.

Dalenius, T., & Reiss, S. P. (1982). Data-swapping: A technique for disclosure control. *Journal of Statistical Planning and Inference*, *6*, 73-85.

Domingo-Ferrer, J. (2002). *Privacy in statistical databases* (LNCS 2316). Springer. Retrieved from http://www.springer.com/computer/theoretical+computer+science/book/978-3-540-43614-0

Domingo-Ferrer, J., & Franconi, L. (2006). *Privacy in statistical databases* (LNCS 4302). Springer. Retrieved from http://www.springer.com/computer/database+management+%26+information+retrieval/book/978-3-540-49330-3?changeHeader

Domingo-Ferrer, J., & Magkos, E. (2010). *Privacy in statistical databases* (LNCS 6344). Springer. Retrieved from http://www.springer.com/computer/database+management+%26+information+retrieval/book/978-3-642-15837-7?changeHeader

Domingo-Ferrer, J., & Saygin, Y. (2008). *Privacy in statistical databases* (LNCS 5262). Springer. Retrieved from http://www.springer.com/computer/database+management+%26+information+retrieval/book/978-3-540-87470-6?changeHeader

Domingo-Ferrer, J., Sebe, F., & Castella-Roca, J. (2004). On the security of noise addition for privacy in statistical databases. In J. Domingo-Ferrer & V. Torra (Eds.), *Privacy in statistical databases 2004* (LNCS 3050) (pp. 149-161). Verlag Berlin Heidelberg: Springer.

Domingo-Ferrer, J., & Torra, V. (2003). Disclosure risk assessment in statistical microdata protection via advanced record linkage. *Statistics and Computing*, *13*, 343-354.

Domingo-Ferrer, J., & Torra, V. (2004). *Privacy in statistical databases* (LNCS 3050). Springer. Retrieved from http://www.springer.com/computer/database+management+%26+information+retrieval/book/978-3-540-22118-0?changeHeader

Duncan, G. T., Elliot, M., & Salazar-González, J. (2011). Statistical confidentiality: Principles and practice. New York, NY: Springer.

Duncan, G. T., Jabine, T. B., & de Wolf, V. A. (1993). *Private lives and public policies*. Washington, DC: National Academy Press.

Dwork, C. (2006). Differential privacy. In *33rd International Colloquium on Automata, Languages and Programming, part II*. Venice, Italy: Springer-Verlag.

Dwork, C., & Naor, M. (2010). On the difficulties of disclosure prevention in statistical database or the case for differential privacy. *Journal of Privacy and Confidentiality*, *2*, 93-107.

El Emam, K. (2011). Methods for the de-identification of electronic health records for genomic research. *Genome Medicine*, *3*, 1-9.

El Emam, K., Brown, A., AbdelMalik, P., Neisa, A., Walker, M., Bottomley, J., & Roffey, T. (2010). A method for managing re-identification risk from small geographic areas in Canada. *BMC Medical Informatics & Decision Making*, *10*, 18.

El Emam, K., Kamal Dankar, F., Issa, R., Jonker, E., Amyot, D., Cogo, E., & Bottomley, J. (2009). A globally optimal k-anonymity method for the de-identification of health data. *Journal of the American Medical Informatics Association*, *16*, 670-682.

Elliot, M. J. (2000). DIS: A new approach to the measurement of statistical disclosure risk. *International Journal of Risk Management*, *2*, 39-48.

Fayyoumi, E., & Oommen, B. J. (2010). A surveyon statistical disclosure control and micro-aggregation techniques for secure statistical databases. *Software: Practice & Experience*, *40*, 1161-1188.

Fienberg, S., & McIntyre, J. (2005). Data swapping: Variations on a theme by Dalenius and Reiss. *Journal of Official Statistics*, *21*, 309-323.

Fuller, W. A. (1993). Masking procedures for microdata disclosure limitation. *Journal of Official Statistics*, *9*, 383-406.

Fung, B. C. M., Wang, K., Chen, R., & Yu, P. S. (2010). Privacy-preserving data publishing: A survey of recent developments. *ACM Computing Surveys*, *42*, 14:1-14:53.

Gouweleeuw, J. M., Kooiman, P., Willenborg, L. C. R.J., & De Wolf, P.-P. (1998). Post randomisation for statistical disclosure control: Theory and implementation. *Journal of Official Statistics*, *14*, 463-478.

Hundepool, A., Wetering, A., Ramaswamy, R., Franconi, L., Polettini, S., Capobianchi, A., & Giessing, S. (2008). *Mu-Argus, Version 4.2 User's Manual*. The Hague, Netherlands: Statistics. Retrieved from http://neon.vb.cbs.nl/casc/Software/MuManual4.2.pdf

Kennickell, A., & Lane, J. (2007). *Measuring the impact of data protection techniques on data utility: Evidence from the survey of consumer finances* (SCF Working Papers). Retrieved from http://www.federalreserve.gov/pubs/oss/oss2/papers/Disclosure3.pdf

Kinney, S. K., Karr, A. F., & Gonzalez, J. F. (2009). Data confidentiality: the next five years summary and guide to papers. *Journal of Privacy and Confidentiality*, *1*, 125-134.

Lane, J. (2007). Optimizing the use of microdata: an overview of the issues. *Journal of Official Statistics*, *23*, 299-317.

Lane, J., Heus, P., & Mulcahy, T. (2008). Data access in a cyber world: Making use of cyberinfrastructure. *Transactions on Data Privacy*, *1*, 2-16.

Lane, J., & Schur, C. (2010). Balancing access to health data and privacy: A review of the issues and approaches for the future. *Health Services Research*, *45*, 1456-1467.

Little, R. J. A., & Liu, F. (2003). Comparison of SMIKe with data-swapping and PRAM for statistical disclosure control of simulated microdata. In *Proceedings of the Section on Survey Research Methods, CD-ROM*. American Statistical Association. Retrieved from http://www.amstat.org/sections/srms/Proceedings/

Loukides, G., Denny, J., & Malin, B. (2010). The disclosure of diagnosis codes can breach research participants' privacy. *Journal of the American Medical Association*, *17*, 322-327.

Malin, B., Sweeney, L., & Newton, E. (2003). Trail re-identification: learning who you are from where you have been. *Workshop on Privacy in Data, Carnegie Mellon University, Pittsburgh, PA*.

McCallum, A., & Wellner, B. (2003). Object consolidation by graph partitioning with a conditionally-trained distance metric. In *Proceedings of the ACM Workshop on Data Cleaning, Record Linkage and Object Identification*, Washington, DC.

Nadeau, C., Gagnon, E., & Latouche, M. (1999). *Disclosure control strategy for the release of microdata in the Canadian Survey of Labour and Income Dynamics*. Paper presented at the 1999 Joint Statistical Meetings, Baltimore, MD.

Narayanan, A., & Shmatikov, V. (2008). Robust de-anonymization of large sparse datasets. In *Proceedings of the 2008 IEEE Symposium on Security and Privacy*. Retrieved from http://www.cs.utexas.edu/~shmat/

Narayanan, A., & Shmatikov, V. (2010). Myths and fallacies of "personally identifiable information." *Communications of the ACM*, *53*, 24-26.

Ochoa, S., Rasmussen, J., Robson, C., & Salib, M. (2008). Reidentification of individuals in Chicago's homicide database—A technical and legal study. Retrieved from http://web.mit.edu/sem083/www/assignments/reidentification.html

Raghunathan, T. E., Reiter, J. P., & Rubin, D. B. (2003). Multiple imputation for statistical disclosure limitation. *Journal of Official Statistics*, *19*, 461-468.

Rastogi, V., Suciu, D., & Hong, S. (2007, September 23-27). The boundary between privacy and utility in data publishing. In *Proceedings of the 33rd International Conference on Very Large Data Bases*. Vienna, Austria.

Reiter, J. (2003). Inference for partially synthetic, public use microdata sets. *Survey Methodology*, *29*, 181-188.

Reiter, J. (2009). Multiple imputation for disclosure limitation: Future research challenges. *Journal of Privacy and Confidentiality*, *1*, 223-233.

Rosenbaum, S. (2010). Data governance and stewardship: designing data stewardship entities and advancing data access. *Health Services Research*, *45*, 1442-1455.

Rothstein, M. (2010). Is deidentification sufficient to protect health privacy in research? *American Journal of Bioethics*, *10*, 3-11.

Rubin, D. B. (1993). Statistical disclosure limitation. *Journal of Official Statistics*, *9*, 461-468.

Scheuren, F., & Winkler, W. (1997). Regression analysis of data files that are computer matched—Part II. *Survey Methodology*, 23, 157-165.

Singh, A. (2009). Maintaining analytic utility while protecting confidentiality of survey and nonsurvey data. *Journal of Privacy and Confidentiality*, *1*, 155-182.

Singh, A., Yu, F., & Dunteman, G. (2003, April). *MASSC: A new data mask for limiting statistical information loss and disclosure* (Working Paper No. 23). Paper presented at the Joint ECE/Eurostat work session on statistical confidentiality, Luxemburg. Retrieved from http://www.unece.org/fileadmin/DAM/stats/documents/2003/04/confidentiality/wp.23.s.e.pdf

Skinner, C. J. (2007). The probability of identification: Applying ideas from forensic statistics to disclosure risk assessment. *Journal of the Royal Statistical Society. Series A: Statistics in Society*, *170*, 195-212.

Skinner, C. J. (2009). Statistical disclosure control for survey data. In D. Pfeffermann & C. R. Rao (Eds.), *Handbook of statistics* (Vol. 29a, pp. 381-396). Amsterdam, Netherlands: Elsevier. Retrieved from http://www.elsevier.com/wps/find/bookdescription.cws_home/719334/description#description

Skinner, C. J., & Shlomo, N. (2008). Assessing identification risk in survey microdata using log-linear models. *Journal of the American Statistical Association*, *103*, 989-1001.

Sweeney, L. (1997). Weaving technology and policy together to maintain confidentiality. *Journal of Law, Medicine & Ethics*, *25*, 98-110.

Sweeney, L. (2000). *Uniqueness of simple demographics in the U.S. population* (LIDAP-WP4). Laboratory for International Data Privacy, Carnegie Mellon University, Pittsburgh, PA.

Sweeney, L. (2002). k-Anonymity: A model for protecting privacy. *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems*, *10*, 557-570.

Sweeney, L. (2010a, March 8-9). *Data sharing under HIPAA: 12 years later*. Paper Presented at HHS Workshop on the HIPAA Privacy Rule's De-identification Standar, Washington, DC. Retrieved from http://hhshipaaprivacy.com/assets/5/resources/Panel2_Sweeney.pdf

Sweeney, L. (2010b). *Designing a Trustworthy Nationwide Health Information Network (NHIN) Promises Americans Privacy and Utility, Rather Than Falsely Choosing Between Privacy or Utility* (Testimony before the 21st Century Healthcare Caucus Round Table, U.S. Congress April 22, 2010). Retrieved from http://patientprivacyrights.org/wp-content/uploads/2010/04/Sweeney-CongressTestimony-4-22-10.pdf

Truta, T. M., Fotouhi, F., & Barth-Jones, D. (2004). Assessing global disclosure risk in masked Microdata. In *Proceedings of the 2004 Workshop on Privacy in Electronic Society* (pp. 85-93). Washington, DC.

United Nations. (2007). *Principles and guidelines for managing statistical confidentiality and microdata access*. Retrieved from http://unstats.un.org/unsd/statcom/doc07/BG-Microdata-E.pdf

Weinberg, D. H., Abowd, J. M., Steel, P. M., Zayatz, L., & Rowland, S. K. (2007). *Access methods for United States microdata* (U.S. Census Bureau Center for Economic Studies Paper No. CES-WP-07-25). Retrieved from http://ssrn.com/abstract=1015374

Willenborg, L., & de Waal, T. (1996). *Statistical disclosure control in practice, lecture notes in statistics*. New York, NY: Springer-Verlag.

Winkler, W. E. (1997). *Views on the production and use of confidential microdata* (Statistical Research Division report RR 97/01). Retrieved from http://www.census.gov/srd/www/byyear.html

Winkler, W. E. (2004a). Masking and re-identification methods for public-use microdata: Overview and research problems. In J. Domingo-Ferrer & V. Torra (Eds.), *Privacy in statistical database* (pp. 231-247). New York, NY: Springer.

Winkler, W. E. (2004b). Re-identification methods for masked microdata. In J. Domingo-Ferrer & V. Torra (Eds.), *Privacy in statistical databases* (pp. 216-230). New York, NY: Springer.

Winkler, W. E. (2007). *Examples of easy-to-implement, widely used methods of masking for which analytic properties are not justified* (Research Report Series #2007-21). Statistical Research Division, U.S. Census Bureau. Retrieved from http://www.census.gov/srd/papers/pdf/rrs2007-21.pdf

Bios

Sergio I. Prada, PhD in public policy, is a research associate at IMPAQ International LLC. His areas of expertise are program evaluation and analytical methods, microeconometrics, economic policy analysis, and health policy statistics.

Claudia A. González Martínez, Ph.D. in Economics, is a senior research associate at Optimal Solutions Group, LLC. Her research interests focus on Health and Labor Economics in general, and in particular on program evaluation, empirical applications and policy analysis.

Josh Borton has been Survey Statistician at NORC since May 2010. His focus for the past 18 months has been statistical disclosure limitation of health data, working with both Electronic Medical Records and Medicare Claims. Josh's diverse experience includes industrial research developing high temperature materials processing technology, and marketing analytics consulting where he applied survival analysis to the valuation of subscription based products.

Johannes Fernandes-Huessy, B.A., is a research analyst in the Economics, Labor, and Population Studies division at NORC at the University of Chicago. Johannes supports the NORC Data Enclave team as the lead Enclave manager. His primary research focus is on illicit drug markets.

Craig Holden, MBA, MPH, CHES, is a senior research analyst at NORC at the University of Chicago. His background includes experience in health research, care delivery, business management and identification and use of performance metrics in program evaluation.

Elizabeth C. Hair, Ph.D., is a senior research scientist in the Public Health Area at NORC at the University of Chicago. She has 20 years of experience in conducting research on child and family well-being. Her expertise includes performance measurement, evaluation, technical assistance, de-identified data, questionnaire design, and advanced statistical modeling.

Timothy M. Mulcahy, is a senior research scientist with 18 years of experience in social science research developing and implementing complex, data-centric projects involving sensitive data, evidence-based research, and data warehousing. He has published and served as an invited speaker, keynote, panel chair, and panelist at numerous conferences, workshops, and seminars issues related to data security, cyber infrastructure, data access modalities, privacy, confidentiality, and statistical disclosure control.