Published in final edited form as:

ACM J Data Inf Qual. 2016 October ; 7(4): . doi:10.1145/2956554.

Preserving Patient Privacy When Sharing Same-Disease Data

XIAOPING LIU*,

Department of Operations and Information Systems, University of Massachusetts Lowell, Lowell, MA 01854; Xiaoping_Liu@student.uml.edu

XIAO-BAI LI*,

Department of Operations and Information Systems, University of Massachusetts Lowell, Lowell, MA 01854

LUVAI MOTIWALLA,

Department of Operations and Information Systems, University of Massachusetts Lowell, Lowell, MA 01854; Luvai_Motiwalla@uml.edu

WENJUN LI,

Department of Medicine, University of Massachusetts Medical School, Worcester, MA 01655; Wenjun.Li@umassmed.edu

HUA ZHENG, and

Department of Orthopedics and Physical Rehabilitation, University of Massachusetts Medical School, Worcester, MA 01655; Hua.Zheng@umassmed.edu

PATRICIA D. FRANKLIN

Department of Orthopedics and Physical Rehabilitation, University of Massachusetts Medical School, Worcester, MA 01655; Patricia.Franklin@umassmed.edu.

Abstract

Medical and health data are often collected for studying a specific disease. For such same-disease microdata, a privacy disclosure occurs as long as an individual is known to be in the microdata.

Individuals in same-disease microdata are thus subject to higher disclosure risk than those in microdata with different diseases. This important problem has been overlooked in data-privacy research and practice, and no prior study has addressed this problem. In this study, we analyze the

---

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies show this notice on the first page or initial screen of a display along with the full citation. Copyrights for components of this work owned by others than ACM must be honored.

Abstracting with credit is permitted. To copy otherwise, to republish, to post on servers, to redistribute to lists, or to use any component of this work in other works requires prior specific permission and/or a fee. Permissions may be requested from Publications Dept., ACM, Inc., 2 Penn Plaza, Suite 701, New York, NY 10121-0701 USA, fax +1 (212) 869-0481, or permissions@acm.org.

Authors' addresses: X.-B. Li (corresponding author), Xiaobai_Li@uml.edu.

*Xiaoping Liu and Xiao-Bai Li are co-first authors.

ACM Reference Format:

Xiaoping Liu, Xiao-Bai Li, Luvai Motiwalla, Wenjun Li, Hua Zheng, and Patricia D. Franklin. 2016. Preserving patient privacy when sharing same-disease data. J. Data Information Quality 7, 4, Article 17 (September 2016), 14 pages.

DOI: <http://dx.doi.org/10.1145/2956554>

CCS Concepts: • Information systems → Data analytics; • Social and professional topics → Medical information policy; Patient privacy

disclosure risk for the individuals in same-disease microdata and propose a new metric that is appropriate for measuring disclosure risk in this situation. An efficient algorithm is designed and implemented for anonymizing same-disease data to minimize the disclosure risk while keeping data utility as good as possible. An experimental study was conducted on real patient and population data. Experimental results show that traditional reidentification risk measures underestimate the actual disclosure risk for the individuals in same-disease microdata and demonstrate that the proposed approach is very effective in reducing the actual risk for same-disease data. This study suggests that privacy protection policy and practice for sharing medical and health data should consider not only the individuals' identifying attributes but also the health and disease information contained in the data. It is recommended that data-sharing entities employ a statistical approach, instead of the HIPAA's Safe Harbor policy, when sharing same-disease microdata.

Keywords

Data sharing; disclosure risk; HIPAA

1. INTRODUCTION

Health data sharing has greatly facilitated medical research, as well as health care delivery. Secondary use of data collected originally for patient care or research can substantially expand knowledge about diseases and treatments, improve quality of health care, support public health policies, and reduce cost of data collection [Basole et al. 2015]. There is clearly a trend that health care organizations are increasingly sharing their health and patient data for secondary use. Furthermore, organizations and researchers are often required to share data collected in publicly funded research. Since 2003, the National Institutes of Health (NIH) has established a data-sharing policy for NIH-sponsored research, stating that, "data should be made as widely and freely available as possible while safeguarding the privacy of participants" [NIH 2003]. The National Science Foundation (NSF) also started a data-sharing policy in 2011, requiring all proposals to include a data-management plan concerning the dissemination and sharing of research results [NSF 2011]. While data sharing has significantly enhanced the quality and efficiency of medical and health care research, there is a growing concern about privacy due to such use of health data. The privacy concern is not limited to health and medical domains. In today's data-driven economy, sharing customer and consumer data is becoming a common practice in business, but privacy concerns have weakened the accessibility and quality of the data for sharing. Therefore, information privacy is an important issue for data-quality research [Christen et al. 2014].

When sharing patient health data, organizations and researchers are required to comply with the privacy guidance specified in the Health Insurance Portability and Accountability Act (HIPAA). HIPAA delineates two approaches for protecting individually identifiable health information [Department of Health and Human Services (DHHS) 2000; DHHS 2002]. The *Safe Harbor* (SH) rule specifies 18 categories of explicitly or potentially identifying attributes, called Protected Health Information (PHI), that must be removed or altered before the health data is released to a third party. Most of the 18 PHI categories are direct

identifiers, such as name, phone number, and e-mail address. There are two PHI categories that are not direct identifiers: dates (e.g., date of birth) and locations (e.g., ZIP code). The SH rule requires that all date values be curtailed to include the year only and ZIP code values be truncated to show the first 3 digits only. To reduce information loss caused by the SH-based deidentification, HIPAA also provides the guidelines for releasing a Limited Data Set (LDS), which contains some date and location information more detailed than that specified under the SH rule. LDS requires data use agreements between the parties involved [DHHS 2000; DHHS 2002].

As an alternative to the SH rule, HIPAA also delineates a *Statistical Standard* approach that enables a statistical assessment of disclosure risk to determine if the data is appropriate for release. A well-known privacy model along this line of approach is k -anonymity [Sweeney 2002]. The k -anonymity model focuses on a type of attribute, called *quasi-identifier* (QI), which includes the date and location attributes considered in SH, as well as other demographic attributes such as age and gender. The values of the QI attributes can often be obtained from public sources, which can be used to reidentify individuals in the deidentified data. To reduce reidentification risk, k -anonymity generalizes the values of QI attributes such that the values of these attributes for any individual matches those of at least $k - 1$ other individuals in the same microdata. In this way, the individual identities are expected to be better protected.

An important task in taking a statistical approach is to analyze privacy disclosure risks. There are two types of disclosure typically recognized in the literature [Duncan and Lambert 1989]: (a) *identity disclosure* or *reidentification*, which occurs when an adversary is able to match a record in a deidentified dataset to an actual individual; and (b) *attribute disclosure*, which occurs when an adversary is able to predict the sensitive value(s) of an individual record, with or without knowing the identity of the individual. The k -anonymity aims to protect against identity disclosure by ensuring the QI values of any individual to be indistinguishable from those of at least $k-1$ other individuals. However, if these k individuals have the same sensitive attribute value (e.g., a disease), then the adversary can achieve attribute disclosure, that is, disclosing the sensitive value of the target individual even though the individual is not definitely identified (because the individual has the same QI values as those of at least $k-1$ other individuals).

In medical and health research, data are often collected for studying a specific disease. In this situation, it is quite likely that all the patients in the entire dataset have the same disease. We call such data *same-disease microdata*. Even though this microdata may also include individuals who do not have the disease (e.g., for comparison purposes), the records with and without the disease are typically known when the data are shared for secondary use. Same-disease microdata is common in medical and health research; examples include cancer registry [Centers for Disease Control and Prevention (CDC) 1992], diabetes cohort studies [van Dam et al. 2006], and registry of HIV patients [Rabeneck et al. 2001]. In its data-sharing policy guidance document, NIH [2003] provides three examples of data-sharing plans. Two are related to same-disease cases. For same-disease microdata, a privacy disclosure occurs as long as an individual is known to be in the microdata (e.g., HIV registry), even though the individual cannot be identified. Thus, individuals in same-disease

microdata are subject to higher disclosure risk than those with different diseases. In considering disclosure risk, neither the Safe Harbor nor statistical/computational approach (e.g., k -anonymity) has differentiated same-disease data from those having different diseases. Therefore, it is necessary to establish an appropriate disclosure risk metric for same-disease microdata. This disclosure is different from identity disclosure or attribute disclosure described earlier. To formally study this disclosure-risk problem, we call the presence of an individual in a microdata set an *instance* and the disclosure of such a presence (without identifying the matching record for the individual) an *instance disclosure*.

In this study, we perform an instance disclosure-risk analysis for same-disease microdata and develop an effective approach to anonymizing the data adequately. The main contributions of this article include the following: (1) We show that Safe Harbor underestimates the disclosure risk for same-disease microdata and k -anonymity provides a misinformed risk estimate that can cause anonymized data to be either underprotected or overprotected. (2) We propose a new disclosure-risk measure and develop an efficient algorithm based on the proposed risk measure for anonymizing same-disease microdata. (3) Using two real patient datasets, we demonstrate the effectiveness of the proposed approach. Furthermore, we provide insights that are valuable to policymakers, data-sharing entities, and data-quality researchers and practitioners.

The rest of the article is organized as follows. In Section 2, we provide background information and related work. The details of the proposed approach are presented in Section 3, in which we discuss reidentification risks with HIPAA and k -anonymity, define an instance disclosure-risk measure for sharing same-disease microdata, and develop an efficient algorithm for anonymizing same-disease data accordingly. Section 4 describes an experimental study that compares our approach with the HIPAA Safe Harbor approach using two real-world datasets. Section 5 discusses the policy and practical implications of this study. We present our conclusions in Section 6.

2. BACKGROUND AND RELATED WORK

2.1. Data Quality Assessment in Privacy-Preserving Data Sharing

For general business settings, a number of quality measures have been established in the literature for assessing data and information quality [Pipino et al. 2002; Madnick et al. 2009]. These measures are, in general, compatible and consistent with respect to a common goal of providing valuable information. Data-quality problems in privacy-preserving data sharing are fundamentally different from traditional data quality problems in that there are two rather inconsistent objectives in the data-sharing scenario. The first objective is to minimize privacy-disclosure risks in the shared data in order to protect privacy. This means that the identifying and sensitive information must be either removed or altered before data is released, which causes information loss and reduces the utility of shared data. Consequently, the second objective is to keep the information loss as small as possible and to maintain the utility of the data for sharing. As a result, for a privacy-preserving data-sharing problem, there are two types of “data-quality” metrics, one for measuring disclosure risk and the other for measuring information loss or data utility.

Disclosure risk is typically measured by the probability of identifying an individual in the released data or the probability of finding the sensitive value(s) of an individual with or without knowing the identity of the individual [Adam and Wortmann 1989]. Disclosure-risk measures are related to the free-of-error and completeness measures in the general data-quality literature [Pipino et al. 2002], but in an opposite sense (e.g., a high error rate in identifying an individual implies a small disclosure risk, thus is desirable). These measures often take a simple ratio or percentage form. For example, the maximum disclosure risk for an individual in a k -anonymized dataset is $1/k$ (to be explained in detail later).

Various solution approaches have been proposed to reduce disclosure risks. In health data-privacy practice, perhaps the most commonly used approach is *generalization*, which generalizes the original values to a higher-level category (Sweeney 2002; Garfinkel et al. 2007; Li and Sarkar 2014). This is exactly the approach taken by k -anonymity. The other main approaches include *noise-based perturbation*, which adds noise to the sensitive data to disguise their true values (Liew et al. 1985; Li and Sarkar 2013) and *data swapping*, which involves exchange of attribute values between different records (Dalenius and Reiss 1982; Li and Sarkar 2011). This work focuses on the generalization approach.

Information loss or data utility measures are directly related to the free-of-error, completeness, and relevancy measures in the data-quality literature [Pipino et al. 2002]. Their actual definitions may depend on data-anonymization methods used. For instance, when k -anonymity is applied, the original dataset is divided into a number of subsets, each containing the records that share the same QI attribute values. The number of subsets is an information-loss measure. A small number indicates that the QI values of the original records are generalized to a few very high-level categories, suggesting a large information loss, while a large number suggests a lower degree of generalization, thus a small information loss. Data-utility measures may also depend on application context. For example, if the anonymized data is to be used for building a prediction model, then the prediction accuracy based on the anonymized data will be an important data-utility measure. We should point out that, when evaluating the effectiveness of different privacy-preserving data-sharing approaches, it is important to examine the performances on both disclosure-risk and data-utility measures.

2.2. Related Work on Privacy Disclosure Risk

The essential idea behind the HIPAA policy is to protect patient privacy against identity disclosure; HIPAA does not provide guidelines on how to protect attribute disclosure. Following HIPAA, data-privacy studies in medical and health care domains focus mostly on identity disclosure. Several studies have considered reidentification risks in the context of population data [Sweeney 2000; Golle 2006]. It was estimated that somewhere between 63% [Golle 2006] and 87% [Sweeney 2002] of the US population can be uniquely identified with three QI attributes: gender, date of birth, and 5-digit ZIP code. Because these studies focus exclusively on population data and do not consider any patient microdata, it remains unclear how the population reidentification risks relate to patient reidentification risks when sharing patient microdata.

Several other studies have examined reidentification risk in microdata [LeFevre et al. 2006; Li and Sarkar 2011]. These studies attempted to estimate reidentification risks based on both microdata and population data. Due to the difficulty in obtaining real population data, the studies typically used “surrogate” populations, such as random samples of a population, certain segments of a population, and summarized tables of census data. Statistical methods were then applied to estimate reidentification risks. However, these studies do not consider attribute disclosure risk.

Attribute disclosure problems have been studied quite extensively in the privacy literature outside the health data-privacy domain [Duncan and Lambert 1989; Machanavajjhala et al. 2006; Li et al. 2007; Li and Sarkar 2009; Li and Sarkar 2014], for which it is typically assumed that there are multiple sensitive attribute values (e.g., multiple diseases) in microdata. Popular privacy models such as t -diversity [Machanavajjhala et al. 2006] and t -closeness [Li et al. 2007] have been developed to handle various attribute-disclosure problems. However, these models rely on the multiple-sensitive-value assumption to reduce attribute disclosure risk. The main idea is to anonymize data such that sensitive attribute values are well distributed for the individuals having the same QI attribute values. When the sensitive attribute has only a single value, as in the same-disease case, none of these approaches is applicable. As mentioned earlier, for same-disease microdata, a privacy disclosure occurs whenever an individual is known to be in the microdata. This disclosure, which we have called instance disclosure earlier, is different from identity disclosure or multivalued attribute disclosure described earlier.

Although same-disease data sharing is quite common, its privacy implications have been overlooked in data-privacy research. Prior approaches in the literature typically assume that there are different diseases in a released dataset. Same-disease data sharing calls for a focused study and a new approach to more effectively deal with its specific privacy-risk problem. In this work, we analyze disclosure risks for same-disease data and perform an experimental study. We contrast the instance-disclosure risk with the disclosure risks implied by the HIPAA SH rule and k -anonymity. We define the instance-disclosure measure and provide real estimates of instance-disclosure risk using real population data and patient microdata.

3. THE PROPOSED APPROACH

3.1. Reidentification Risks with HIPAA and k -Anonymity

HIPAA considers identity disclosure based on population data [DHHS 2000; DHHS 2002]. To illustrate the idea, consider an example segment of population data in Table I, which is publicly available (e.g., from voter registration lists). The original data contains two QI (also PHI) attributes: 5-digit ZIP code (Zip5) and date of birth (DOB). The last two columns show their Safe Harbor representation: 3-digit zip code (Zip3) and year of birth (YOB). In data-privacy literature, the set of all records that share the same values on a set of QI attributes is called an *equivalence class* (EC) [LeFevre et al. 2006]. For example, the last two records in the original data, Helen and Irene, form an EC and every other record is an EC individually. With Zip3 and YOB representation, there are only two ECs (separated by the dashed line), one including the first three records and the other containing the remaining six records.

Let N_i be the number of records in the i th EC in population data P . Under Safe Harbor, the *reidentification risk* for each record in the i th EC is

$$q_i = \frac{1}{N_i}.$$

Thus, with original data, the reidentification risk is 1/2 for Helen and Irene and is one (100%) for the other individuals. With Zip3 and YOB representation, the risk is 1/3 for each of the first three records and 1/6 for each of the remaining six records.

For an individual in a microdata set, the reidentification risk is the chance of correctly matching this individual to an individual in the population. This can be calculated based on q_i . Table II shows a patient microdata set in the same format as that of Table I except that the direct identifier, Name, is removed and replaced by a system-generated noninformative Patient ID. If the data is released with Zip5 and DOB, then the first five records can be uniquely reidentified based on the population data: they are Alice, Bob, Charlie, Dave, and Grace, respectively. The last record has a reidentification risk of 1/2 (either Helen or Irene). If the data is released with Zip3 and YOB, then Patient #1 can be Alice, Bob, or Charlie; thus, the reidentification risk for the patient is 1/3. Similarly, the reidentification risk for Patients #2 and #3 is also 1/3, respectively. For Patient #4 (or #5 or #6), there are 6 matching records in the population. Thus, the reidentification risk for Patient #4 (or #5 or #6) is 1/6.

The k -anonymity model does not provide a precise estimate of reidentification risk for an individual record. Instead, it provides the *maximum* reidentification risk for any individual record in a dataset, which is $1/k$. This maximum occurs when the individuals in an EC in the microdata are the same as those in the corresponding EC in the population. When releasing data in Table II, if Zip3 and YOB are used, then the released data satisfy 3-anonymity and the maximum reidentification risk is 1/3 for any record. This maximum risk is equal to the actual reidentification risk for the first three records but much larger than the actual risk (1/6) for the last three records.

3.2. Instance-Disclosure Risk for Same-Disease Microdata

For same-disease data, disclosure risk should be evaluated differently. To see this, assume that all records in the example have the same disease. Suppose that the data is released with Zip3 and YOB, which satisfies both Safe Harbor and 3-anonymity requirements. An adversary having access to the population data will know for certain that the first three records are Alice, Bob, and Charlie. If the adversary's target is Alice (or any of these three people), the adversary will discover that Alice has the disease even though it cannot be determined which of the three patients is Alice. The actual identification of Alice is not important here. Because the number of records in this EC is 3 in both the microdata and the population, the chance of the instance that an individual in the population appears in the microdata is $3/3 = 1$. In terms of the second EC (with Zip3 = '002**' and YOB = 1935), the number of records is 3 in the microdata and 6 in the population. Therefore, the chance of the instance is $3/6 = 0.5$.

Based on this observation, we now define the instance-disclosure=risk. Let D be a same-disease microdata set for which all direct identifiers are removed. Let P be the population segment containing D . In P , direct identifiers exist and the QI attributes are represented in the same way as in D . Thus, for each EC in D , there is an EC in P with the same QI values. We arrange matching ECs in D and P in the same order and label the matching EC in D and P with the same index i . Let n_i and N_i be the number of records in the i th EC in D and P , respectively. The *instance-disclosure risk* for a record in the i th EC is defined by

$$r_i = \frac{n_i}{N_i}.$$

Statistically, r_i is the probability that an individual having the QI values specified in the i th EC in population P appears in microdata D . The instance-disclosure risk has the following important property.

Proposition 1.—Instance-disclosure risk is generally greater than reidentification risk; that is, $r_i \geq q_i$ for every i .

Proof. Since—Since $n_i \geq 1$, we have that

$$r_i = \frac{n_i}{N_i} \geq \frac{1}{N_i} = q_i.$$

It follows from Proposition 1 that the widely used reidentification risk measure actually underestimates the disclosure risk for same-disease data. The maximum reidentification risk suggested by k -anonymity, which is $1/k$, may also underestimate the disclosure risk for same-disease data. This is true for the illustrative example in Table II, in which instance-disclosure risks for the two ECs are 1 and 0.5, both greater than $1/3$. It is also possible for k -anonymity to overestimate the risk for same-disease microdata. Suppose that there are 15 individuals in the population having Zip3 = ‘002**’ and YOB = 1935. Then, the instance-disclosure risk for a record in the second EC in the microdata is $3/15 = 0.2$, which is much smaller than $1/3$. In short, the maximum reidentification risk suggested by k -anonymity does not really provide appropriate information about disclosure risk for same-disease data.

The instance-disclosure risk is defined for an individual record. To measure average risk with respect to a microdata set, let $|D|$ be the number of records in D and m be the number of ECs in D . Then, the *average instance-disclosure risk* for D is defined by

$$R = \frac{1}{|D|} \sum_{i=1}^m n_i r_i.$$

For the illustrative example in Table II, the average instance-disclosure risk is

$$R = \frac{1}{6} [3(1) + 3(0.5)] = 0.75.$$

Like a probability measure, the instance-disclosure risk measure r_i and average instance-disclosure risk R range between zero and one, as stated in Proposition 2 below.

Proposition 2—The values of r_i and R are in the range of $(0, 1]$.

Proof—Clearly, $r_i > 0$, $\forall i$, and $R > 0$, since $n_i, N_i > 0, i \forall i$. It is also obvious that $n_i \leq N_i$ since the number of records in the i th EC in D cannot be greater than the corresponding number in P . Thus, $r_i = n_i/N_i \leq 1$, and

$$R = \sum_{i=1}^m \frac{n_i r_i}{|D|} \leq \sum_{i=1}^m \frac{n_i (1)}{|D|} = \frac{|D|}{|D|} = 1.$$

We can similarly define the average reidentification risk for D as

$$Q = \frac{1}{|D|} \sum_{i=1}^m n_i q_i.$$

For the illustrative example, the average reidentification risk is

$$Q = \frac{1}{6} \left[3 \left(\frac{1}{3} \right) + 3 \left(\frac{1}{6} \right) \right] = 0.25.$$

3.3. The Proposed Algorithm to Reduce Instance-Disclosure Risk

To anonymize the data with the same sensitive value (e.g., same disease), we use the *generalization* operation as in k -anonymity [Sweeney 2002], which generalizes or truncates QI attribute values to higher-level values gradually. In particular, ZIP code values are generalized by removing a digit gradually from right to left. DOB values are first generalized to YOB values and may be further generalized to a range of YOB values (e.g., ‘1935-1940’) if necessary. The following property provides a theoretical basis for our proposed algorithm to anonymize the data using generalization.

Proposition 3—Any generalization of QI attribute values causes the maximum instance-disclosure risk to decrease.

Proof—Without loss of generality, consider two ECs, D_1 and D_2 , which are to be generalized to one EC, $D_0 = D_1 \cup D_2$. Let P_0 , P_1 , and P_2 be the sets of all records that have the same QI attribute values as those in D_0 , D_1 , and D_2 , respectively. In general, P_0 includes not only P_1 and P_2 , but also additional records in the population whose QI values before generalization are different from those of D_1 and D_2 . (For example, suppose that ZIP codes for D_1 and D_2 are 10011 and 10012, respectively. If the generalized ZIP code for D_0 is 1001*, then P_0 will include all records with ZIP code values from 10010 to 10019, some of which may not appear in D_1 or D_2 .) Let P_a be such additional records in P . Then, $P_0 = P_1 \cup P_2 \cup P_a$.

Let instance-disclosure risk for D_0 , D_1 , and D_2 be $r(D_0)$, $r(D_1)$, and $r(D_2)$, respectively. We show that

$$r(D_0) < \max\{r(D_1), r(D_2)\}.$$

Without loss of generality, assume that $r(D_1) \geq r(D_2)$. Then, if $r(D_0) < r(D_1)$, the above result is obtained immediately. Now, consider $r(D_0) \geq r(D_1)$. In this case, we can show that $r(D_0) < r(D_2)$ using proof by contradiction. Assume that this is not true; that is $r(D_0) \geq r(D_2)$. Then, we have that $r(D_0) \geq r(D_1) \geq r(D_2)$. It follows from the definitions $r(D_1) = |D_1| / |P_1|$ and $r(D_2) = |D_2| / |P_2|$ that

$$|D_1| \leq |P_1| r(D_0), \text{ and } |D_2| \leq |P_2| r(D_0).$$

Thus,

$$|D_1| + |D_2| \leq r(D_0) (|P_1| + |P_2|).$$

Substituting the definition of $r(D_0)$ into this inequality, we have that

$$\frac{|D_1| + |D_2|}{|P_1| + |P_2|} \leq r(D_0) = \frac{|D_0|}{|P_0|} = \frac{|D_1| + |D_2|}{|P_1| + |P_2| + |P_a|},$$

which is not true in general. Therefore, either $r(D_0) < r(D_1)$ or $r(D_0) < r(D_2)$. This completes the proof.

An algorithm using generalization to reduce the instance-disclosure risk should be able to consider both microdata and population data. Existing k -anonymity algorithms (e.g., Sweeney [2002] and LeFevre et al. [2006]) are not appropriate because they are based on microdata only. On the other hand, approaches based on reidentification risk are also not applicable because they consider population data only. We propose a novel algorithm that efficiently computes instance-disclosure risks using both microdata and population data. The algorithm divides the data into a number of subsets based on the idea of recursive binary partitioning in decision trees [Breiman et al. 1984; Li and Sarkar 2009, 2014]. Unlike traditional bottom-up k -anonymity algorithms, such as Sweeney [2002], the proposed algorithm divides the data from top down, which is computationally more efficient.

After the dataset is partitioned into subsets, the records within a subset are more similar to each other than those between subsets. For example, the first three records in Table II will most likely be grouped in one subset; the remaining three records will be in another subset. The values of the QI attributes (Zip5 and DOB) for the records within each subset are very similar. These values are then generalized to transform each subset to an EC. To avoid unnecessary information loss, the generalization is based on the most detailed common QI values within a subset. For example, the original ZIP code values for the two subsets in Table II will not be generalized to Zip3 (Safe Harbor) format, but will remain in Zip5 format since all records within the same subset have the same Zip5 value (i.e., 00101 and 00202 for

each subset, respectively). On the other hand, DOB will be generalized to YOB. It is clear that, in general, the utility of the data processed using the proposed algorithm is expected to be better than that based on the SH rule.

In the recursive partitioning process, there are many ways to split the data by using different QI attributes and different values of a QI attribute, causing different instance-disclosure risks and different data qualities when the QI values of the partitioned subsets are generalized. We have discussed how to measure instance-disclosure risk. In terms of data quality, it is clear that an attribute having a larger variance in its values will have more information loss if the values of the attribute are generalized. Such an attribute should have a higher priority to be selected for partition to reduce the variance after partition. Let v_j be the variance of attribute j in a (partitioned) dataset. Let $R_{\chi(s)}$ be the average instance-disclosure risk when splitting the subset at value s of attribute j . Then, the ratio $R_{\chi(s)}/v_j$ captures both the disclosure risk and data-quality aspects. Because a small disclosure risk and a large variance are preferred, the split having the minimum $R_{\chi(s)}/v_j$ should be selected for partitioning the current set. Our proposed algorithm uses this criterion at each iteration. Note that in computing variance, we first transform categorical QI values into numeric or ordered values based on coding methods suggested in LeFevre et al. [2006], then normalize all original or transformed numeric values to unit scale.

The proposed algorithm is given in Figure 1. It follows from Proposition 3 that the maximum instance-disclosure risk increases as recursive partitioning of dataset D causes the partitioned subsets to be progressively smaller, thus generalized at more detailed levels. Therefore, a minimum subset size, like the k parameter in k -anonymity, can be used to control the disclosure risk. The algorithm is computationally analogous to a decision-tree algorithm. As such, the time complexity of the algorithm is of $O(\text{Mog}N)$, where N is the number of records in P . In actual implementation, we can reduce P to include only the segment of the population that is relevant to D . As such, P is unlikely to be overly large. Thus, the algorithm can be quite efficient. The algorithm assumes that the QI attribute values can be ordered. Otherwise, local recoding, as suggested in LeFevre et al. [2006], should be applied to convert the data to orderable values.

4. EXPERIMENTAL EVALUATION

We conducted an experimental study in order to validate the analytical results obtained in the previous section and to compare the proposed algorithm with the SH approach. The study used two real patient datasets with real population data and was approved by the Institutional Review Board of the authors' institutions. The first dataset includes 180 records of patients who had undergone the same surgical procedure (thus can be considered as same-disease data) in an NIH-funded, single-center, randomized trial listed on the ClinicalTrials.gov website. All patients resided in a single northeast state in the United States, and were recruited between 2008 and 2010. There are three QI attributes in the dataset: gender, date of birth, and 5-digit ZIP code (LDS). The patients were 61% female, had a mean age of 65 years, and resided in 84 ZIP codes. The second dataset includes 300 randomly selected records of patients who resided in the same state as the first dataset and went to the same center in 2012. The QI attributes include gender, year of birth, and 3-digit

ZIP code (SH compliant). The voter registration lists for that state were collected to serve as the primary population data. The full voter dataset included 3,641,990 records. Out of these records, 401,517 were in the 84 ZIP codes, which were used in the experiment. A commercial data vendor was also used as a supplemental source for population data. The population data includes detailed information for the individuals, such as gender, residence address, and date of birth.

We first compare the results of reidentification risk with those of instance-disclosure risk for SH and LDS release. Note that LDS is applicable to the first dataset only since the second set was originally in SH format already. As described earlier, reidentification risk and instance-disclosure risk vary with different records. Thus, we report in Table III the maximum reidentification risk ($\max q_i$) and maximum instance-disclosure risk ($\max r_i$), as well as the average risks Q and R . It is clear from Table III that $\max q_i$ and Q are considerably smaller than $\max r_i$ and R , respectively, in all scenarios (except the maximum risks in LDS release). This suggests that traditional reidentification risk measures seriously underestimate the real risk of disclosure for same-disease data. It is also observed that LDS release has much higher risks than SH release for all risk measures, which is expected. Both $\max q_i$ and $\max r_i$ with LDS release are one (100%), indicating unique reidentification of at least one record in the dataset.

Next, we examine the effectiveness of the proposed algorithm in reducing instance-disclosure risk in comparison with the SH approach. This was performed for the first dataset only because the ZIP code and date attributes in the second dataset were originally given in Zip3 and YOB format, which do not provide room for the proposed algorithm to generalize the values at a more detailed level. We anonymize the original data with the SH rule and the proposed algorithm, respectively. In generalizing QI values for a dataset, it is clear that the larger the number of ECs (i.e., the smaller the size of an EC), the less degree of generalization is required for individual ECs, which means less information loss after generalization. To facilitate the comparison, we have thus used our algorithm to partition the data such that the number of ECs generalized is no less than the number of ECs with the SH approach, which implies that information loss for the data generalized with our algorithm is no more than that with the SH approach.

The results from SH and the proposed algorithm are shown in Table IV. The number of ECs with the proposed algorithm is slightly larger than that with SH, suggesting slightly smaller information loss with our algorithm. On the other hand, the maximum and average instance-disclosure risks with our algorithm are only about half of those with SH. Therefore, our algorithm is very effective in reducing instance-disclosure risk for same-disease data. We have also reported runtime performance in Table IV. Apparently, it takes more time for the proposed algorithm to anonymize data than the SH approach. This is expected because, as discussed earlier, the time complexity for the proposed algorithm is of $O(\text{Mog}N)$, while it is linear for the SH approach. Nevertheless, the proposed algorithm is efficient and fast enough for practical use.

5. DISCUSSION

Sharing of same-disease data is common in medical research and health care practice. Individuals in same-disease microdata are subject to higher disclosure risk than those in microdata with different diseases. This problem has been overlooked in data-privacy research and practice. In this study, we have shown analytically and experimentally that the widely used reidentification risk measure underestimates the actual disclosure risk for same-disease data. With increasing concerns for patient privacy, this finding has significant policy and practical implications.

This study reveals two limitations of the HIPAA Safe Harbor policy. First, Safe Harbor applies the same standards for releasing different microdata, which expectedly causes underprotection for some microdata but overprotection for others because disclosure risks in different microdata are different. In the same-disease case, Safe Harbor tends to be underprotective. Second, SH considers only PHI elements, based exclusively on identity-disclosure concerns. Studies have shown that sensitive data that are not PHI, such as disease data [Machanavajjhala et al. 2006; Li et al. 2007] and mobility traces [de Montjoye et al. 2013], can cause privacy breach. Instance disclosure in same-disease data poses another kind of privacy threat not caused by identity disclosure. This suggests that focusing on PHI alone without considering disease information may not be adequate for safeguarding patient privacy. Same-disease data requires tighter privacy protection than data with different diseases.

However, we do not advocate setting up a more restrictive Safe Harbor standard. A more stringent Safe Harbor policy would cause overly large information loss for many data-sharing applications. We recommend that data-owner organizations and researchers employ HIPAA's Statistical Standard approach when sharing same-disease microdata. This work has established a theoretical ground for such a statistical approach. As shown in this article, disclosure risk analysis for same-disease microdata is, in a sense, simpler than that for data with different diseases (such as l -diversity and t -closeness). It is worthwhile to make the effort to pursue the analysis.

For the broad data-quality research community, it is important to note that, for a privacy-preserving data-sharing problem, there are two types of data-quality measures with inconsistent objectives: privacy disclosure risk and data utility. When evaluating different privacy-preserving data-sharing approaches, both disclosure-risk and data-utility measures must be examined together.

6. CONCLUSION

We have performed an analytical and experimental study of the disclosure risk for same-disease microdata. In closing, we should emphasize that privacy implications vary across different diseases. For example, an HIV patient dataset is obviously much more sensitive than a flu patient dataset. Therefore, even though the flu dataset has higher disclosure risks than the HIV dataset, it is expected that the HIV dataset requires a more protective action. This should be very clear to policymakers and data-sharing entities. In addition, we should

stress that the proposed approach applies to the same-disease data-sharing problem but is inappropriate for more general multiple-disease problems.

In this study, we have focused on instance-disclosure risk and data utility without any application context. However, data utility can be measured in various ways, depending on application context. Different data utility measures can be used for same-disease data analytics. For example, if the purpose is to build a classification model based on the data to help determine if a patient has a disease, then classification error as well as false-positive and false-negative errors will be essential data-utility measures [Li and Sarkar 2009]. On the other hand, if the task is to discover association rules among health data items, then an important data utility measure should be accuracy or error rate in the support of large itemsets [Li 2009].

One limitation of this work is that the proposed algorithm was tested on only one real dataset for proof of concept. This dataset might not be an ideal representation of various patient populations. For example, there are more female than male and much older patients in the data, and all patients resided in a single state in the United States. It will be more helpful if more datasets with different characteristics are used for experimental evaluation. In order to compare the proposed algorithm with the Safe Harbor approach, the PHI values in the original data need to be more detailed than those restricted by Safe Harbor (e.g., date of birth instead of year of birth, and 5-digit ZIP code instead of 3-digit ZIP code). Due to data holders' privacy concerns, it is very difficult to obtain data with more detailed information than that allowed by Safe Harbor. Future research will obtain more data to further validate the proposed approach.

Acknowledgments

This research was supported by the National Library of Medicine (NLM) and National Institute of Arthritis and Musculoskeletal and Skin Diseases (NIAMS) of the National Institutes of Health (NIH) under Grant Numbers R01LM010942 and R01AR054479. The content is solely the responsibility of the authors and does not necessarily represent the official views of NLM, NIAMS, or NIH.

REFERENCES

- Adam NR, Wortmann JC. Security-control methods for statistical databases: A comparative study. *ACM Computing Surveys*. 1989; 21(4):515–556.
- Basole RC, Braunstein ML, Sun J. Data and analytics challenges for a learning healthcare system. *ACM Journal of Data and Information Quality*. 2015; 6:2–3. Article 10, 4.
- Breiman, L.; Friedman, JH.; Olshen, RA.; Stone, CJ. *Classification and Regression Trees*. Wadsworth; Belmont, CA.: 1984.
- Christen P, Vatsalan D, Verykios VS. Challenges for privacy preservation in data integration. *ACM Journal of Data and Information Quality*. 2014; 5:1–2. Article 4, 3.
- Centers for Disease Control and Prevention (CDC). [August 11, 2016] National Program of Cancer Registries. 1992. from <http://www.cdc.gov/cancer/npcr/about.htm>
- Dalenius T, Reiss SP. Data swapping: A technique for disclosure control. *Journal of Statistical Planning and Inference*. 1982; 6(1):73–85.
- de Montjoye YA, Hidalgo CA, Verleysen M, Blondel VD. Unique in the crowd: The privacy bounds of human mobility. *Scientific Reports*. 2013; 3 Article 1376.
- Department of Health and Human Services (DHHS). Standards for privacy of individually identifiable health information. *Federal Register*. 2000; 65(250):82462–82829.

- Department of Health and Human Services (DHHS). Standards for privacy of individually identifiable health information. Federal Register. 2002; 67(157):53181–53273. [PubMed: 12180470]
- Duncan GT, Lambert D. The risk of disclosure for microdata. *Journal of Business and Economic Statistics*. 1989; 7(2):201–217.
- Garfinkel R, Gopal R, Thompson S. Releasing individually identifiable microdata with privacy protection against stochastic threat: An application to health information. *Information Systems Research*. 2007; 18(1):23–41.
- Golle, P. Proceedings of the 5th ACM Workshop on Privacy in Electronic Society (WPES'06). ACM; New York, NY: 2006. Revisiting the uniqueness of simple demographics in the US population.; p. 77–80.
- LeFevre, K.; DeWitt, DJ.; Ramakrishnan, R. Mondrian multidimensional k-anonymity. In Proceedings of the 22nd International Conference on Data Engineering (ICDE'06).. IEEE Computer Society; Washington, DC. 2006. p. 25–35.
- Li, N.; Li, T.; Venkatasubramanian, S. t-Closeness: Privacy beyond k-anonymity and l-diversity. In Proceedings of the 23rd IEEE International Conference on Data Engineering (ICDE'07).. IEEE Computer Society; Washington, DC. 2007. p. 106–115.
- Li XB. A Bayesian approach for estimating and replacing missing categorical data. *ACM Journal of Data and Information Quality*. 2009; 1(1) Article 3, 11.
- Li XB, Sarkar S. Against classification attacks: A decision tree pruning approach to privacy protection in data mining. *Operations Research*. 2009; 57(6):1496–1509.
- Li XB, Sarkar S. Protecting privacy against record linkage disclosure: A bounded swapping approach for numeric data. *Information Systems Research*. 2011; 22(4):774–789.
- Li XB, Sarkar S. Class-restricted clustering and microperturbation for data privacy. *Management Science*. 2013; 59(4):796–812.
- Li XB, Sarkar S. Digression and value concatenation to enable privacy-preserving regression. *MIS Quarterly*. 2014; 38(3):679–698. [PubMed: 26752802]
- Liew CK, Choi UJ, Liew CJ. A data distortion by probability distribution. *ACM Transactions on Database Systems*. 1985; 10(3):395–411.
- Machanavajjhala, A.; Gehrke, J.; Kifer, D.; Venkatasubramanian, M. Proceedings of the 22nd IEEE International Conference on Data Engineering. (ICDE'06). IEEE Computer Society; Washington, DC: 2006. l-Diversity: privacy beyond k-anonymity.; p. 24–35.
- Madnick SE, Lee YW, Wang RY, Zhu H. Overview and framework for data and information quality research. *ACM Journal of Data and Information Quality*. 2009; 1(1) Article 2, 22.
- National Institutes of Health (NIH). [August 11, 2016] NIH Data Sharing Policy and Implementation Guidance. 2003. from http://grants.nih.gov/grants/policy/data_sharing/data_sharing_guidance.htm
- National Science Foundation (NSF). [August 11, 2016] Dissemination and Sharing of Research Results. 2011. from <http://www.nsf.gov/bfa/dias/policy/dmp.jsp>
- Pipino LL, Lee YW, Wang RY. Data quality assessment. *Communications of the ACM*. 2002; 45(4): 211–218.
- Rabeneck L, Menke T, Simberkoff MS, Hartigan PM, Dickinson GM, Jensen PC, George WL, Goetz MB, Wray NP. Using the national registry of HIV-infected veterans in research: Lessons for the development of disease registries. *Journal of Clinical Epidemiology*. 2001; 54(12):1195–1203. [PubMed: 11750188]
- Sweeney, L. Uniqueness of simple demographics in the U.S. population. Working paper, LIDAP-WP4. Data Privacy Lab, Carnegie Mellon University; Pittsburgh, PA.: 2000.
- Sweeney, L. *International Journal of Uncertainty*. Vol. 10. Fuzziness and Knowledge-Based Systems; 2002. k-Anonymity: A model for protecting privacy.; p. 557–570.
- van Dam RM, Willett WC, Manson JE, Hu FB. Coffee, caffeine, and risk of type 2 diabetes: A prospective cohort study in younger and middle-aged U.S. women. *Diabetes Care*. 2006; 29(2): 398–403. [PubMed: 16443894]

-
0. Given: microdata D and underlying population P , which have d common QI attributes.
  1. List the values of each QI attribute in ascending or descending order. With the values ordered, trial splits can be performed linearly between every pair of adjacent values.
  2. For each QI attribute j , compute $R_{j(s)}/v_j$ for each trial split s . Call the trial split having the minimum $R_{j(s)}/v_j$ the “best trial split” for QI attribute j .
  3. Find the overall best split among the d best trial splits and partition the current dataset into two subsets using the attribute value of the overall best split.
  4. Repeat Steps 2 and 3 for each of the two subsets until a prespecified stopping criterion is met (e.g., the minimum number of records required for a subset).
  5. For each subset, generalize the QI attribute values using their most detailed common value within the subset, which transforms the subset into an EC.
-

Fig. 1.

Algorithm to generalize data based on instance-disclosure risk.

Table I

An Illustrative Example of Population Data

Name	5-Digit ZIP Code (Zip5)	Data of Birth (DOB)	3-Digit ZIP Code (Zip3)	Year of Birth (YOB)
Alice	00101	07/15/1927	001**	1927
Bob	00101	05/28/1927	001**	1927
Charlie	00101	10/26/1927	001**	1927
Dave	00202	01/02/1935	002**	1935
Emily	00202	02/03/1935	002**	1935
Frank	00202	10/24/1935	002**	1935
Grace	00202	05/13/1935	002**	1935
Helen	00202	09/26/1935	002**	1935
Irene	00202	09/26/1935	002**	1935

Table II

An Illustrative Example of Same-Disease Microdata

Patient ID	5-Digit ZIP Code (Zip5)	Data of Birth (DOB)	3-Digit ZIP Code (Zip3)	Year of Birth (YOB)
1	00101	07/15/1927	001**	1927
2	00101	05/28/1927	001**	1927
3	00101	10/26/1927	001**	1927
4	00202	01/02/1935	002**	1935
5	00202	05/13/1935	002**	1935
6	00202	09/26/1935	002**	1935

Table III

Results of Reidentification Risks and Instance-Disclosure Risks

Data Release Method	Dataset Name	$\max q_i$	$\max r_i$	Q	R
Safe Harbor (SH)	Dataset 1	0.0376	0.0473	0.0020	0.0043
	Dataset 2	0.0382	0.0474	0.0027	0.0042
Limited Dataset (LDS)	Dataset 1	1.0000	1.0000	0.5344	0.6633

Table IV

Results from Safe Harbor and Proposed Algorithm on Dataset 1

Data Release Method	Number of ECs	$\max r_i$	R	Runtime
Safe Harbor	109	0.0256	0.0030	0.5s
Proposed Algorithm	120	0.0154	0.0016	7.0s