Privacy Preservation in Releasing Patient Data

Full papers

***Xiaoping Liu, Xiao-Bai Li, Luvai Motiwalla**
Department of Operations and Information Systems
University of Massachusetts Lowell
*Email: xiaoping_liu@uml.student.edu

Wenjun Li, Hua Zheng, Patricia D. Franklin
University of Massachusetts Medical School

Abstract

When patient data are shared for studying a specific disease, a privacy disclosure occurs as long as an individual is known to be in the shared data. Individuals in such specific disease data are thus subject to higher disclosure risk than those in datasets with different diseases. This problem has been overlooked in privacy research and practice. In this study, we analyze disclosure risks for this problem and identify appropriate risk measures. An efficient algorithm is developed for anonymizing the data. An experimental study is conducted to demonstrate the effectiveness of the proposed approach.

Keywords

Privacy, Patient Data, Risk

Introduction

When patient data are shared for medical research and healthcare practice, it is required that appropriate measures be taken to protect the privacy of the patients. HIPAA (Health Insurance Portability and Accountability Act) delineates two approaches for protecting individually identifiable health information (DHHS 2000). 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 email 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 de-identification, 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.

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 attributes, called *quasi-identifier* (QI), which include 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 re-identify individuals in the de-identified data. To reduce re-identification risk, *k*-anonymity generalizes the values of QI attributes such that the values of these attributes for any individual match those of at least $k-1$ other individuals in the same microdata. In this way, the individual identities are expected to be better protected.

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 the *same-disease microdata*. Even though the 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. The same-disease microdata is common in medical and health research; examples include cancer registry, diabetes cohort studies, and registry of HIV patients. For the 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 the same-disease microdata are subject to higher disclosure risk than those with different diseases. In considering disclosure risk, SH and well-known statistical approaches (e.g., k-anonymity) do not differentiate the same-disease data from those having different diseases. Therefore, it is necessary to establish an appropriate method for evaluating disclosure risk for the same-disease microdata.

In this study, we perform a disclosure risk analysis for the same-disease microdata and develop an effective approach to anonymizing the data adequately. We show that Safe Harbor underestimates the disclosure risk for the same-disease microdata and k-anonymity provides misinformed risk estimate that can cause the anonymized data to be either under-protected or over-protected. We developed an efficient algorithm for anonymizing the same-disease data. Using a real patient datasets, we demonstrate the effectiveness of the proposed approach.

Background and Related Work

In analyzing privacy disclosure risk, the literature typically recognizes two types of disclosure (Duncan and Lambert 1989): (a) *identity disclosure* or *re-identification*, which occurs when an adversary is able to match a record in a de-identified 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, i.e., disclosing the sensitive value of the target individual even though the individual is not definitely identified. Similar to k-anonymity, HIPAA also focuses on identity disclosure problem and does not provide guidelines on how to protect attribute disclosure.

Following HIPAA, data privacy studies in medical and healthcare domains focus mostly on identity disclosure. Several studies have considered re-identification risks in the context of population data (Benitez and Malin 2010; Sweeney 2002), while others have examined re-identification risk in microdata (El Emam et al. 2011). These studies, however, do not consider attribute disclosure risk.

Attribute disclosure problems have been studied quite extensively in the privacy literature outside health privacy domain (Machanavajjhala et al. 2006; Li et al. 2007), where it is typically assumed that there are multiple sensitive attribute values in microdata. Popular privacy models such as l-diversity (Machanavajjhala et al. 2006) and t-closeness (Li et al. 2007) have been developed to handle various attribute disclosure problems. These models, however, 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 the same-disease microdata, a privacy disclosure occurs as long as an individual is known to be in the microdata. This disclosure is different from identity disclosure or multi-valued attribute disclosure described above. To formally study this disclosure problem, we call the presence of an individual in a microdata set an *instance* and the disclosure of such a presence an *instance disclosure*.

Re-identification Risks with HIPAA and k-Anonymity

HIPAA considers identity disclosure based on population data (Benitez and Malin 2010; El Emam et al. 2011). To illustrate the idea, consider an example segment of population data in Table 1, 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 SH 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). In the original data, for example, the last two records, 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 dash-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 *re-identification risk* for each record in the *i*th EC is

$$q_i = 1/N_i$$

So, with original data, the re-identification 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 re-identification risk is the chance of correctly matching this individual to an individual in the population. This can be calculated based on q_i. Table 2 shows a patient microdata set in the same format as that of Table 1 except that the direct identifier, Name, is removed and replaced by a system generated non-informative Patient ID. If the data are released with Zip5 and DOB, then the first five records can be uniquely re-identified based on the population data – they are Alice, Bob, Charlie, Dave and Grace, respectively. The last record has a re-identification 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; so the re-identification risk for the patient is 1/3. Similarly, the re-identification 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. So, the re-identification risk for patient #4 (or #5 or #6) is 1/6.

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 1. An Illustrative Example of Population Data

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 2. An Illustrative Example of the Same-Disease Microdata

The k-anonymity model does not provide a precise estimate of re-identification risk for an individual record. Instead, it provides the maximum re-identification risk for any 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 2, if Zip3 and YOB are used, then the released data satisfy 3-anonymity and the maximum re-identification risk is 1/3 for any record. This maximum risk is equal to the actual re-identification risk for the first three records but much larger than the actual risk (1/6) for the last three records.

Instance Disclosure Risk for the Same-Disease Microdata

For the same-disease data, disclosure risk should be evaluated differently. To see this, assume all records in the above example have the same disease. Suppose the data is released with Zip3 and YOB, which satisfies both SH and 3-anonymity requirements. An adversary having an access to the population data will know for certain that the first three records are Alice, Bob and Charlie. If his target is Alice (or any of these three people), he will discover that Alice has the disease even though he cannot determine 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, 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 the above observation, we now define the instance disclosure risk. Let *D* be a same-disease microdata set where all the direct identifiers are removed. Let *P* be the population segment containing *D*. In *P* direct identifiers exists and the QI attributes are represented in the same way as in *D*. So, 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

$$r_i = n_i/N_i$$

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*. Comparing r_i with q_i, since $n_i > 1$ in general, it is clear that instance disclosure risk is generally greater than re-identification risk. Therefore, the widely used re-identification risk measure actually underestimates the disclosure risk for the same-disease data. The maximum re-identification risk suggested by *k*-anonymity, which is $1/k$, may also underestimate the disclosure risk for the same-disease data. This is true for the illustrative example in Table 2, where 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 the same-disease microdata. Suppose 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 re-identification risk suggested by *k*-anonymity does not really provide appropriate information about disclosure risk for the 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 2, the average instance disclosure risk is

$$R = [3(1) + 3(0.5)]/6 = 0.75$$

We can similarly define the average re-identification risk for *D* as

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

For the illustrative example, the average re-identification risk is

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

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.

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; LeFevre et al. 2006) are not appropriate because they are based on microdata only. On the other hand, approaches based on re-identification risk are also not applicable because they consider population data only. We propose an algorithm to reduce the instance disclosure risk using both microdata and population data. It divides the data into a number of subsets based on the idea of recursive partitioning in decision trees. The QI attribute values in the subsets 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 zip code values for the two subsets in Table 2 will not be generalized to Zip3 format but will remain in Zip5 format since all records within the same subset have the same Zip5 value (i.e., 00101 and 00202 respectively). On the other hand, DOB will be generalized to YOB.

In the recursive partitioning process, there are many ways to split the data by using different QI attribute values, 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 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 $R_{j(s)}$ be the average instance disclosure risk when splitting the subset at value s of attribute j. Then, the ratio $R_{j(s)}/v_j$ 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 $R_{j(s)}/v_j$ 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], and then normalize all original or transformed numeric values to unit scale.

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 $R_{j(s)}/v_j$ as the "best trail-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 pre-specified 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

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 and thus be generalized at more detailed levels. So, 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 $O(N\log N)$, 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. So, 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.

Experimental Evaluation

We conducted experiments on a real patient dataset, which includes 180 records of patients who had undergone the same surgical procedure (and thus can be considered as the same-disease data). All patients resided in a single northeast state in the US. 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 voter registration lists for that state were collected to serve as the primary population data. A commercial data vendor was also used as a supplemental source for population data.

Release Method	max q_i	max r_i	Q	R
Safe Harbor	0.0376	0.0473	0.0020	0.0043
Limited Data Set	1.0000	1.0000	0.5344	0.6633

Table 3. Results of Re-identification Risks and Instance Disclosure Risks

We first compare the results of re-identification risk with those of instance disclosure risk for SH and LDS release. As described earlier, re-identification risk and instance disclosure risk vary with different records. So, we report in Table 3 the maximum re-identification 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 3 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 re-identification risk measures seriously underestimate the real risk of disclosure for the 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 re-identification of at least one record in the dataset.

Next, we examine the effectiveness of the proposed algorithm in reducing the instance disclosure risk in comparison with the SH approach. 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.

Release Method	Number of ECs	max r_i	R
Safe Harbor	109	0.0256	0.0030
Proposed Algorithm	115	0.0137	0.0015

Table 4. Results from Safe Harbor and Proposed Algorithm on Dataset 1

The results from SH and the proposed algorithm are shown in Table 4. The number of ECs with the proposed algorithm is slightly larger than that with SH, suggesting slightly smaller information loss with our algorithm. The maximum and average instance disclosure risks with our algorithm, on the other hand, are only about half of those with SH. Therefore, our algorithm is very effective in reducing instance disclosure risk for the same-disease data.

Discussion

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

This study reveals two limitations of the SH policy. First, SH applies the same standards for releasing different microdata, which expectedly causes under-protection for some microdata but over-protection for others because disclosure risks in different microdata are different. In the same-disease case, SH tends to be under-protective. Second, SH considers only PHI elements, based exclusively on identity disclosure concern. Instance disclosure in the same-disease data poses a 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. The same-disease data require tighter privacy protection than data with different diseases.

We do not, however, advocate setting up a more restrictive SH standard. A more stringent SH 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 the same-disease microdata. This work has established a theoretical ground for such a statistical approach. As shown in this paper, disclosure risk analysis for the 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 take the effort to pursue the analysis.

Conclusion

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 the policy makers and data-sharing entities.

Acknowledgements

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

References

Benitez, K., and Malin, B. 2010. "Evaluating Re-Identification Risks with Respect to the HIPAA Privacy Rule," *Journal of the American Medical Informatics Association* (17:2), pp. 169-177.

Department of Health and Human Services (DHHS). 2000. "Standards for Privacy of Individually Identifiable Health Information," *Federal Register* (65:250), pp. 82462-82829.

Duncan, G.T., and Lambert, D. 1989. "The Risk of Disclosure for Microdata," *Journal of Business and Economic Statistics* (7:2), pp. 201-217.

El Emam, K., Paton, D., Dankar, F., and Koru, G. 2011. "De-identifying a Public Use Microdata File from the Canadian National Discharge Abstract Database," *BMC Medical Informatics and Decision* Making (11), Article 53, 26 pp.

LeFevre, K., DeWitt, D.J., and Ramakrishnan, R. 2006. "Mondrian Multidimensional *k*-Anonymity," in *Proceedings of the 22nd International Conference on Data Engineering* (ICDE'06), Washington, DC: IEEE Computer Society, pp. 25-35.

Li, N., Li, T., and Venkatasubramanian, S. 2007. *t*-Closeness: Privacy Beyond *k*-Anonymity and *l*-Diversity," in *Proceedings of the 23rd IEEE International Conference on Data Engineering* (ICDE'07), Washington, DC: IEEE Computer Society, pp. 106-115.

Machanavajjhala, A., Gehrke, J., Kifer, D., and Venkitasubramaniam, M. 2006. "l-Diversity: Privacy Beyond k-Anonymity," in Proceedings of the 22nd IEEE International Conference on Data Engineering. (ICDE'06), Washington, DC: IEEE Computer Society, pp. 24-35.

Sweeney, L. 2002. "k-Anonymity: "A Model for Protecting Privacy," International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems (10:5) pp. 557-570.