Evaluating re-identification risks with respect to the HIPAA privacy rule

Kathleen Benitez,[1] Bradley Malin[1,2]

► Supplementary appendices are published online only at http://jamia.bmj.com/content/vol17/issue2

[1]Department of Biomedical Informatics, School of Medicine, Vanderbilt University, Nashville, Tennessee, USA [2]Department of Electrical Engineering and Computer Science, School of Engineering, Vanderbilt University, Nashville, Tennessee, USA

Correspondence to
Bradley Malin, 2525 West End Avenue, Suite 600, Department of Biomedical Informatics, School of Medicine, Vanderbilt University, Nashville, TN 37203, USA; b.malin@vanderbilt.edu

Received 4 April 2009
Accepted 14 December 2009

ABSTRACT

Objective Many healthcare organizations follow data protection policies that specify which patient identifiers must be suppressed to share "de-identified" records. Such policies, however, are often applied without knowledge of the risk of "re-identification". The goals of this work are: (1) to estimate re-identification risk for data sharing policies of the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule; and (2) to evaluate the risk of a specific re-identification attack using voter registration lists.

Measurements We define several risk metrics: (1) expected number of re-identifications; (2) estimated proportion of a population in a group of size g or less, and (3) monetary cost per re-identification. For each US state, we estimate the risk posed to hypothetical datasets, protected by the HIPAA Safe Harbor and Limited Dataset policies by an attacker with full knowledge of patient identifiers and with limited knowledge in the form of voter registries.

Results The percentage of a state's population estimated to be vulnerable to unique re-identification (ie, $g=1$) when protected via Safe Harbor and Limited Datasets ranges from 0.01% to 0.25% and 10% to 60%, respectively. In the voter attack, this number drops for many states, and for some states is 0%, due to the variable availability of voter registries in the real world. We also find that re-identification cost ranges from $0 to $17 000, further confirming risk variability.

Conclusions This work illustrates that blanket protection policies, such as Safe Harbor, leave different organizations vulnerable to re-identification at different rates. It provides justification for locally performed re-identification risk estimates prior to sharing data.

INTRODUCTION

Advances in health information technology have facilitated the collection of large quantities of finely detailed personal data,[1] which, in addition to supporting innovative healthcare operations, has become a vital component of numerous secondary endeavors, including novel comparative quality research and the validation of published findings.[2 3] Historically, data collection and processing efforts were performed internally by the same organization, but sharing data beyond the borders of the organization has become a vital component of emerging biomedical systems.[2 3] In fact, it is of such importance that in the United States, some federal agencies such as the National Institutes of Health (NIH) have adopted policies that mandate sharing data generated or studied with federal funding.[4 5]

To realize the benefits of sharing data while minimizing privacy concerns, many healthcare organizations have turned to "de-identification", a technique that strips explicit identifying information, such as personal names or Social Security Numbers, from disclosed records. Healthcare organizations often employ multiple tiers of de-identification policies, the appropriateness of which is usually dependent on the recipient and intended use. Each policy specifies a set of features that must be suppressed from the data. Presently, healthcare organizations tend to employ at least two policy tiers: (1) *public use*; and (2) *restricted access research*. The public use policy removes a substantial number of explicit identifiers and "quasi-identifying", or potentially identifying, attributes. The resulting dataset is thought to contain records that are sufficiently resistant to privacy threats. In contrast, the restricted access research policy retains more detailed features, such as dates and geocodes. In return for additional information, oversight or explicit approval from the originating organization is required.

Though de-identification is a widely invoked approach to privacy protection, there have been limited investigations into the effectiveness of such policies. Anecdotal evidence suggests that concerns over the strength of such protections may be warranted. In 1996, for instance, Sweeney was able to merge publicly available de-identified hospital discharge records with identified voter registration records on the common fields of *date of birth*, *gender* and *residential zip code* to re-identify the medical record for the governor of Massachusetts, uncovering the reason for a mysterious hospital stay.[6] In subsequent investigations, it was estimated that somewhere between 63% and 87% of the US population is unique on the combination of such demographics.[6 7] However, both investigations assumed that an "attacker" has ready access to a resource with names and demographics for the entire population.

There are several primary goals and contributions of this paper. First, we extend earlier work[6 7] by defining and applying several computational metrics to determine the extent to which de-identification policies in the Privacy Rule of the Health Insurance Portability and Accountability Act[8] (HIPAA) leave populations susceptible to re-identification. In particular, we focus on the Safe Harbor and Limited Dataset policies, which, akin to the policy tiers mentioned earlier, define public use and restricted use datasets. In the process, we illustrate how to compare the re-identification risk tradeoffs between competing policies. We perform this analysis in a generative manner and assume that an

attacker has access to all the identifying information on the de-identified population. Second, we demonstrate how to model concerns in a more realistic setting and consider the context of a limited knowledge attacker. Specifically, while the analysis mentioned in the first part of the paper assumes access to identifying information for the entire population, the accessibility of such data cannot be taken for granted. And, while voter registration lists have been exploited in one known instance and are cited as a source of identified data, such an attack may not be feasible in all situations. We investigate how the real world availability of voter registration resources influences the re-identification risks. Voter information is often managed at the state level, and thus we perform our analysis on a state-by-state basis to determine how blanket federal-level data sharing policies (ie, HIPAA) are affected by regional variability. Our results show that differences in risk are magnified when the wide spread of state voter registration policies is taken into account. Overall, our study provides evidence that the risks vary greatly and an attacker's likelihood of re-identification success is dependent on the population from which the released dataset is drawn.

BACKGROUND

In this section, we review the foundations of de-identification and re-identification. We examine previous privacy risk analysis approaches and illustrate the concepts with a motivating example.

From de-identification to re-identification

Consider the hypothetical situation outlined in figure 1. In this setting, a healthcare provider maintains identified, patient-level clinical information in its private *medical records*. For various reasons, the provider needs to share aspects of this data with a third party, but certain fields in the dataset are sensitive, and therefore an administrator must take steps to protect the privacy of the patients. The de-identification policy of the provider forbids the disclosure of personal names and geographic attributes, so these fields are suppressed to create the *released dataset*. The residual information, however, may still be susceptible to re-identification.

In this work, we are concerned with attacks that re-identify as many records as possible, which in prior publications have been called marketer attacks.[i] A large-scale attack requires an identified dataset having fields in common with the de-identified dataset, such as the fictional *voter list* in figure 1. A re-identification, also known in the literature as an identity disclosure,[9] is accomplished when an attacker can make a likely match between a de-identified record and the corresponding record in the identified dataset. For simplicity, we assume that identified public records contain data on everyone in the de-identified release, making the identified population a superset of the de-identified dataset. We acknowledge this is a simplification and point out that it results in a worst-case risk analysis; that is, an upper bound on the number of possible re-identifications. The online appendix elaborates on this component of the problem.

Unique individuals are most vulnerable to re-identification precisely because matches are certain in the eyes of an attacker. In figure 1, for instance, there is only one person in the population who is a male born in 1953. As a result, since he is a patient in the released dataset, his identity, which is reported in the voter list, can easily be linked to his record in the released dataset. However, it is important for the reader to recognize that uniqueness is only a sufficient, and not a necessary, condition for achieving re-identification. Anytime there is a level of individuality, or *distinctiveness* as we shall call it, there is the potential for re-identification. Notice, again in figure 1, that there are two records in the released dataset for male patients born in 1955. Similarly, there are also two males born in 1955 in the population at large. While these records are non-unique, an attacker who linked the identities to the sensitive records through a random assignment procedure would be correct half of the time.

Identified datasets and the use of voter registration records

The key to successfully achieving a large-scale re-identification attack is the availability of an identified dataset with broad population coverage. In this sense, public records can provide for an easily accessible resource that often includes richly-detailed demographic features. While identified records with features linkable to de-identified data could be obtained through illegitimate means, such as the theft of a laptop that stores such lists on an unencrypted hard drive (eg, see Tennessee[10]) or hacking a state-owned website (eg, see Illinois[11]), lawful avenues make it possible for potential attackers to obtain some public records, such as voter registration lists, without committing any crime. Moreover, access to such records can, in some cases, be obtained without a formally executed data use agreement.

In this paper we focus on voter registration information as a route of potential re-identification for several reasons. First, as mentioned in the introduction of this paper, this resource was applied in one of the most famous re-identification studies to date: the case study by Sweeney.[6] Second, following in the footsteps of this case study, there have been a significant number of publications by the academic and policy communities that suggest such records are a particularly enticing resource for would-be attackers.[12–21] However, allusions to the potential uses of voter lists rarely acknowledge the complexity of data access intricacies, or the economics, of the attack. Rather, they tend to make an implicit assumption that a universal set of demographic attributes tied to personal identity is available to all potential adversaries for a nominal fee. But the reality of the situation is that, if not the absolute contrary, the ability to apply such a resource for re-identification is not universal. Consider, in 2002, a survey of voter registration data gathering and privacy policies which documented that, while all but one state required voters to provide their date of birth, 11 states redacted certain features associated with date of birth prior to making records available to secondary users.[21] The accessibility of identifying resources, such as voter registration lists, is made even more complex by the fact that state-level access policies for identified records are dynamic and change over time. To generate results that are relevant to the current climate, this paper updates the aforementioned survey.

Re-identification risk measures

Most risk evaluation metrics for individual level data focus on one of the following factors: (1) the number, or proportion, of unique individuals; or (2) the worst case scenario, that is, the identifiability of the most vulnerable record in the dataset.

Of those that consider the first factor, the most common approach simply analyzes the proportion of records that are unique within a particular population.[22 23] Alternative approaches that have been proposed add nuance, for instance not just considering unique links, but the probability that a unique link between sensitive and identified datasets is correct. This accounts for the complexities of the relationship between the populations represented (further details on this matter are provided in online Appendix B).[24] The second body of work

[i]For further discussion of the types of attacks and types of re-identifications, see online Appendix A.

Medical Records

ID	Name	Gender	Date of Birth	Hometown	Diagnosis
1	Sister Susie	F	1/1/1953	Lafayette, IN	Myeloid leukemia
2	Jack Sprat	M	3/15/1953	Lafayette, IN	Hypertension
3	Mary Contrary	F	2/28/1953	Washington, IN	Myocardial infarction
4	Boy Blue	M	7/4/1955	Washington, IN	Myocaridal infarction
5	King Cole	M	3/31/1957	Lafayette, IN	Diabetes
6	Jill Hill	F	1/12/1955	Washington, IN	Diabetes
7	Jack Hill	M	1/12/1955	Washington, IN	hypertension

Voter List

Name	Gender	Year of Birth
Sister Susie	F	1953
Jack Sprat	M	1953
Mary Contrary	F	1953
Boy Blue	M	1955
King Cole	M	1957
Jill Hill	F	1955
Jack Hill	M	1955
Jane Goose	F	1957
Jack Nimble	M	1956
Betty Blue	F	1956

Released Dataset

Gender	Year of Birth	Diagnosis
F	1953	Myeloid leukemia
M	1953	Hypertension
F	1953	Myocardial infarction
M	1955	Myocardial infarction
M	1955	Hypertension

Figure 1 Example of de-identification and re-identification using public records.

comes into play when none of the records is likely to be unique.[9] These approaches define disclosure risk as the probability that a re-identification can be achieved.

For the evaluation offered in this paper, we adopt a measure proposed by Truta *et al*,[25] which offers an advantage over the narrow focus on either unique individuals or the most susceptible individuals. This measure incorporates risk estimates for all records in the dataset, regardless of their level of distinctiveness.

METHODS

Materials

We utilized the following resources for our evaluation: (1) HIPAA policies for secondary data sharing to determine the fields available in released datasets; (2) real voter registration access policies for each US state to determine the fields available to an attacker; and (3) demographic summary statistics from the 2000 US Census as population descriptors. We describe each of these resources in the following sections.

Sensitive data policies

Medical and health-related records are considered to contain sensitive information by many people.[26] The unauthorized disclosure of an individual's private health data, such as a positive HIV test result, can have adverse effects on medical insurance, employment, and reputation.[27 28] Yet, health data sharing is vital to further healthcare research, and thus there are various mechanisms for doing so in a de-identified format. As part of HIPAA, for instance, the Privacy Rule regulates the use and disclosure of what is termed "Protected Health Information".[8] Of particular interest to our study are two de-identification policies specified by the Privacy Rule, namely Safe Harbor and Limited Dataset, which permit the dissemination of patient-level records without the need for explicit consent.

The *Safe Harbor* policy enumerates 18 identifiers that must be removed from health data, including personal names, web addresses, and telephone numbers. This process creates a public-use dataset, such that once data has been de-identified under this policy, there are no restrictions on its use. As in many data sharing regulations in the USA and around the world, Safe Harbor contains a special threshold provision for geographic area.[29] When a geographic area (eg, zip code) contains at least 20 000 people, it may be included in Safe Harbor protected datasets, otherwise it must be removed.[ii] Therefore, the threshold of 20 000 is significant for an analysis of population distinctiveness, which we

[ii]For simplicity, we assume no geographic detail beyond "US state" is made available through Safe Harbor.

explicitly investigate in the following evaluation. In contrast, the *Limited Dataset* policy specifies a subset of 16 identifiers that must be removed, creating a research dataset. In order to obtain this dataset, recipients must sign a data use agreement, a contract that restricts the use of the data. Such agreements often explicitly prohibit attempts to re-identify or contact the subjects.

In this paper, we focus explicitly on demographic information, which is particularly relevant to risk analysis because of its wide availability in health and public records, especially in the form of voter registration lists. We assume that an unmodified dataset managed by a healthcare entity includes (*Name*, *Address*, *Date of Birth*, *Gender*, *Race*). When filtered through Safe Harbor, a released dataset will contain only (*Year of Birth*, *Gender*, *Race*), while a Limited Dataset release will also include (*County*, *Date of Birth*).

Voter registration information

Information regarding voter registration lists is available from several sources. Most US state websites maintain online, unofficial versions of their regulatory codes, which contain the policies that govern the use and administration of voter registration lists (eg, Alabama[30]). In some states this information is sufficient to learn which fields are specifically permitted in public releases of the voter registration lists. In other states, the regulations are prohibitory, simply stating which fields cannot be part of the public record. We deemed that a survey of each state's elections office was the most reliable source for information regarding the current contents and prices of voter registration lists. We conducted this survey (results in online Appendix C) in the fall of 2008 by making inquiries with election offices and interpreting a variety of voter registration forms and legal paperwork because there is no standard form or procedure for obtaining state voter lists. Information available in both private health data and voter registration information consists mainly of demographics, such as age, gender, or race.[iii] Thus, we defined the potential fields of intersection as (*Date of Birth*, *Year of Birth*, *Race*, *Gender*, *County of Residence*).

Population information

The census is a natural place to turn for population descriptions subdivided by the aforementioned demographic features. The 2000 US Census is one of the most complete population records to date with an undercount rate estimated to be between 0.96% and 1.4%.[31] Many of the results of the census are freely available online through the Census Bureau's American Fact Finder website.[32] Tables PCT12 A–G detail the number of people of each gender, by age, in a particular geographic division, each table representing one of the Census's seven race classifications: *White alone*, *Black alone*, *American Indian or Alaska Native alone*, *Asian alone*, *Native Hawaiian or Pacific Islander alone*, *Some other race alone*, and *Two or more races*. This information is available for many geographic breakdowns, but as we defined the fields of intersection to include only information as specific as county, the most appropriate division was each table for the 3219 US counties and county equivalents. We created tables for each state and an additional table to translate between field names and the age ranges, genders, and races they represent, so that populations with fields in common could be combined where warranted.

While the census provides the majority of the information needed, it is not a perfect fit. In particular, the census partitions the population by gender and age, whereas voter registration data include year of birth, for which we assume age is a proxy. However, there are additional challenges. For instance, ages over 100 are grouped by the US Census into 5-year age groups (100–104, 105–110). Additionally, information on date of birth is not reported. To overcome such limitations, we leverage a statistical estimation technique proposed by Golle, which is based on the assumption that members of the group are distributed uniformly at random in the larger group.[7] This implies that an individual is as likely to be born on January 5 as January 6, and likewise, that an individual in the age group 100–104 is as likely to be 100 as 101. More generally, given an aggregated group with n individuals who could correspond to b possible subgroups, or "bins", the number of bins with i individuals is estimated as:

$$f_n(i) = \binom{n}{i} b^{1-n}(b-1)^{n-i} \quad (1)$$

As an example, if there are 200 individuals in a group, say 24-year-old "Asian alone" males in County X, then $200 \times 365^{-199} \times 364^{199} \approx 116$ are expected to have a unique birth date.

Risk estimation metrics

We developed two risk estimation metrics that we believe provide a compromise between focusing on likely re-identifications and accepting that there is some probability of re-identification for every record in a released dataset. They are termed g-distinct and *total risk* and are defined as follows.

g-Distinct

An individual is said to be unique when he or she has a combination of characteristics that no one else has, and we say an individual is g-distinct if their combination of characteristics is identical to g-1 or fewer other people in the population. Therefore, uniqueness is the base case of 1-distinct. In general, g-distinct is the sum of the number of bins with i individuals, which is computed as:

$$h_n(g) = \sum_{i=1}^{g} i f_n(i) \quad (2)$$

Of the 200 individuals above, approximately 199.95 would be 5-distinct. It is useful to think of these numbers in terms of proportions rather than absolute numbers. In this case, 99.975% of the group is 5-distinct. Therefore, if a released dataset contained three "Asian only" 24-year-old males, 2.999 of them would be expected to be 5-distinct. Formally, given j members of a group of n, the expected number that will be g-distinct is given as follows:

$$\hat{h}_n^j(g) = \frac{j}{n} h_n(g) \quad (3)$$

Total risk

We extend the notion of g-distinct to cover all possible g's to create a measure of "total risk". This is similar to the DR_{max} metric proposed by Truta *et al*[25] and quantifies the likelihood of re-identification for each member of a group. When summed over all groups, it reveals the expected number of re-identifications for the whole dataset. Specifically, given j members of a group of n, the expected number of re-identifications (ie, the total risk) is computed as:

$$\tilde{r}_n^j(g) = \frac{j}{n} b^{1-n}(b^n - (b-1)^n) \quad (4)$$

Process

The risk analysis estimation consists of a three step process: (1) determine the fields available to an attacker; (2) group the Census data according to these fields; and (3) sum the result obtained by applying a risk estimation metric to the results,

[iii] While voter history is available from many states' voter registration lists, and is not explicitly prohibited by either of the privacy policies under consideration, it is certainly not likely to turn up in a medical record.

normalizing by the total population. The interplay of the data is illustrated in figure 2, which depicts the relationship between our simulation of re-identification (top) and the expected approach of an attacker (bottom).

We consider two types of risk for the purposes of this work, which we call *GENERAL* and *VOTER*. *GENERAL* is the risk associated with a fully informed attacker and corresponds to the worst-case scenario. It assumes that the attacker has access to identifying information for each individual and all the relevant fields for linkage for the entire population from which the disclosed records were derived. To determine the fields available to a *GENERAL* attacker, consider the data protection policy and assume the attacker has access to all the demographic data permitted by that policy. In figure 1, the released dataset has fields (*Gender, Year of Birth, Diagnosis*), so we assume that the attacker has identifying information containing (*Gender, Year of Birth*), and would use these fields to re-identify the released dataset. The *GENERAL* attacker is the typical risk model applied today. The second model, *VOTER*, is tempered in that it considers the availability of a specific identified resource. Specifically, the fields available to a *VOTER* attacker are derived from the data de-identification policy and the voter registration access policy of the relevant state.

Post-analysis calculations

Trust differential

We use the re-identification risk estimates to compare the protective capability of data sharing policies through a mechanism we call the *trust differential*. This term stems from the practice of using several policies to govern the disclosure of the same dataset. In the case of the public and research datasets, the latter contain more information because the researchers are more trusted or are discouraged through various penalties of violating a use agreement. Formally, we model the differential as the ratio of policy-specific risks as $R_{j,g}(A)/R_{j,g}(B)$, where $R_{j,g}(X)$ is the risk measure for the group size g under policy X as computed by re-identification metric j. Imagine that policy A corresponds to *Limited Dataset* and policy B corresponds to *Safe Harbor*. Then, the resulting ratio quantifies the extent to which researchers are more trusted than the general public. Calculation of the trust differential specifies the degree to which the latter policy better protects the data.

Cost analysis

While an economic analysis does not fit strictly into the diagram in figure 1, it is a logical and practical aspect of the voter attack to study. Cost acts as a deterrent in computer security-related incidents,[33] such that an attack on privacy will only be attempted if the net gain is greater than the net cost. Voter registration lists, along with many other identified datasets, may be available to an attacker, but at a certain price. An economic analysis with respect to any of the above measures is then the price in dollars for the resource normalized by the result of the re-identification risk metric, that is C/*R*, where C is the cost for the resource, and *R* is the expected risk to the dataset from an attacker using that resource as computed in equation (4). For example, total risk conveys essentially the expected number of re-identifications. Thus, the economic analysis with respect to total risk will be an estimate of the price the attacker pays for each successful re-identification. All things being equal, we assume an attacker will be more drawn to an attack with a lower cost to success ratio.

RESULTS

For each US state we set g equal to 1, 3, 5, and 10 and for one state, we performed a more detailed analysis, such that g was evaluated over the range 1 through 20 000. We performed a cost analysis using the total risk measure over the same range. For presentation purposes, we have divided the major results of the evaluation to first report results computed with g-distinct, and then results calculated by total risk measures.

In general, we use a combination of factors to perform our risk analysis and use the <*Policy, Attack*> pair to summarize the specific evaluation. *Policy* refers to the health data sharing policy and corresponds to either the Safe Harbor (*SAFE*) or Limited Dataset (*LIMITED*) policy. *Attack* refers to the information we assume is available to the adversary and refers to the *GENERAL* or *VOTER* scenario.

g-Distinct analysis

The g-distinct analysis enables data managers to inspect a particular cross-section of the population, namely the individuals whose records are most vulnerable to re-identification by virtue of being the most distinctive. The plots in figure 3 illustrate the results for the state of Ohio. The analysis of this state is particularly interesting because its voter registration list includes (*County, Year of Birth*) and is thus different from either of the two HIPAA policies. The risk analysis for <*LIMITED, GENERAL*> measures the re-identification risks associated with the Ohio population using the attributes of (*County, Gender, Date of Birth, Race*), and <*LIMITED,*

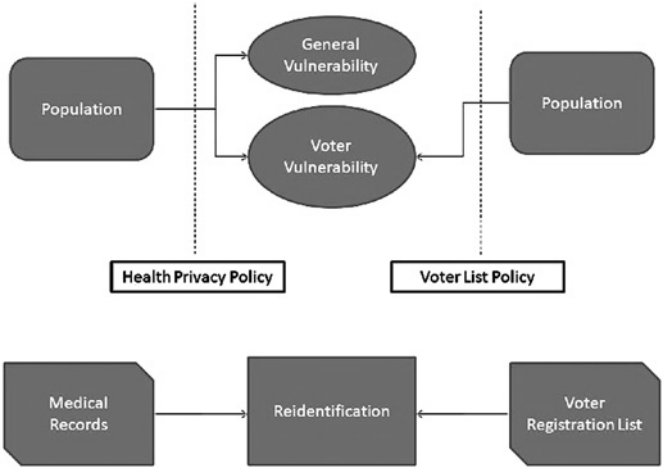

Figure 2 Interplay of data sources in re-identification.

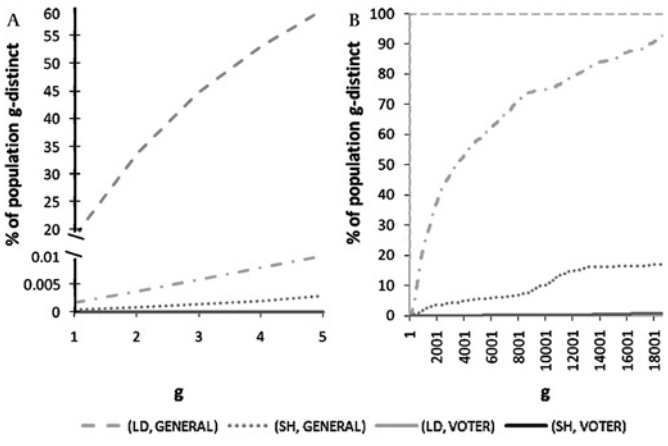

Figure 3 g-Distinct risk analysis for the state of Ohio. (A) g=1 to 5 (B) g=1 to 20 000.

VOTER> using the attributes (*County, Year of Birth*), while the risk analysis for <*SAFE, GENERAL*> uses (*Gender, Year of Birth, Race*), and <*SAFE, VOTER*> uses (*Year of Birth*).

Both plots in figure 3 represent the same result, but at different granularities. The plot on the left focuses on the population that is particularly distinct, those identical to 5 or fewer people. We focus on this cut-off because it is a common risk threshold adopted by many healthcare and statistical agencies. We observe that there is a large gap between the risk associated with Limited Dataset and the other risks measured. Under Limited Dataset, 18.7% of the population is 1-distinct, or unique, and 59.7% are 5-distinct. In contrast, under Safe Harbor, 0.0003% are 1-distinct and 0.002% are 5-distinct. When these patterns are inspected over a wider range of values of g, as shown in the plot on the right, the pattern continues, such that the risk under Limited Dataset rises quickly, surpassing 99.9% by g=31. In other words, fewer than 0.1% of the population in Ohio is expected to share the combination of (*County, Gender, Date of Birth, Race*) with more than 31 people.

The sheer number of distinct individuals can be startling. If a researcher receives a dataset drawn at random from the population of Ohio under Limited Dataset provisions, more than 1 out of 6 of those represented would be unique based on demographic information. Remember, though, that uniqueness is not sufficient to claim re-identification. There is still need for an identified dataset and *VOTER* reflects this reality. While higher than the risk under Safe Harbor, <*LIMITED, VOTER*> is significantly lower than <*LIMITED, GENERAL*> , particularly for smaller values of g. According to <*LIMITED, VOTER*>, only 0.002% of the population is 1-distinct and 0.01% is 5-distinct. As we increase g, we find that more than 50% of the population is 3500-distinct under the same constraints. In other words, very few individuals are readily identifiable with any certainty. In comparison, less than 1% of the population is 20 000-distinct for <*SAFE, VOTER*>. Either way, the probability of re-identification is small, but non-zero.

We can see more precisely how the two policies compare in figure 4, which displays the trust differential for both *GENERAL* and *VOTER*. In *GENERAL*, the trust differential for the two policies ranges from approximately 5 to 90 000, while the *VOTER* trust differential ranges from approximately 67 to more than 3.9 trillion. The extremely high values are found for the lowest values of g, where small differences in values are sufficient to make the differential oscillate, as can be seen in the plot. Consistently, however, the trust differential is large even with g equal to 20 000. It is perhaps an important feature that the trust differential is greatest for low values of g, again, for the individuals who are most susceptible to re-identification.

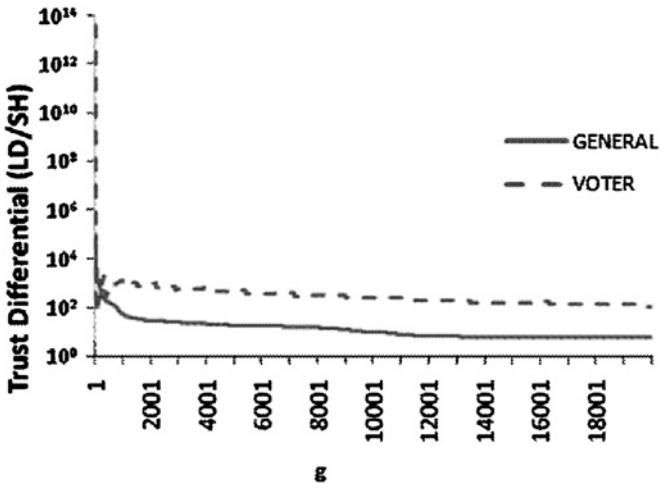

Figure 4 Trust differential (plotted on log scale) between Limited Dataset and Safe Harbor for the state of Ohio.

While the above results demonstrate the power of the g-distinct analysis and the effects of different choices of g, they are not necessarily representative of the results for other states. Thus, figure 5 shows the range of vulnerabilities for selected small values of g for all 50 states (details for all states are in online Appendix D). True to the results found in Ohio, vulnerabilities under Safe Harbor are lower than those under Limited Dataset. Safe Harbor vulnerabilities, however, are spread over a wide range of small values, sufficient to create outliers, seen in both of the Safe Harbor analyses in figure 6. Additionally, notice the reduction of risk when attack-specific information is introduced. While the 10-distinctiveness of the states ranges from 0.44 to nearly 1, with a median of 0.925, the attack-specific 10-distinctiveness ranges from 0 to 0.99, with a median of 0.36. In other words, considering the actual attack tends to much lower risk estimates, particularly when analyzing a less restrictive policy.

Figures 6 and 7 provide another perspective on the results in figure 5. In these plots, we show the two most vulnerable and two least vulnerable states according to 1-distinct, for their respective risk estimate and policy. These results summarize how the state's re-identification risk changes for various g (values for each US state are provided in online Appendix E). Our goal was to characterize how changes in re-identification risk related to each other across states. In other words, we wanted to determine how decisions made for risk thresholds affected the re-identification estimates of the states. For the most part, the rankings remain fairly consistent, but not universally. In particular, we observed that the most substantial change within the range g less than 10 is the state of Kentucky for <*LIMITED, VOTER*>. This state had the second greatest percentage of 1-distinct individuals, but is ninth at the 10-distinct level. Thus, an attacker may shift focus from one state to another depending on the policy and risk threshold.

Total risk analysis

While g-distinct estimates enable analysts to determine which states are the most vulnerable given a particular policy, the total risk measure estimates the number of re-identifications that could theoretically be achieved by an attacker. It is important to recognize that each record has some non-zero probability of being re-identified, even if very small. The total risk measure aggregates these probabilities.

Table 1 displays the results of the total risk analysis for the states with the top three and bottom three trust differentials for *GENERAL* and *VOTER*. A complete list of states and their total risk measures under these policies and types of analysis can be found in online Appendix E. In contrast to the state of Ohio, as previously discussed, the state of Texas's voter registration policy includes all of the fields available in Limited Dataset releases. Therefore, the health record policy is the limiting factor, meaning that *GENERAL* and *VOTER* are identical. For the rest of the states the voter registration policy is the limiting factor, and thus the *GENERAL* and *VOTER* are different. For some states, this is a slight difference, such as Virginia, whereas for others it is several orders of magnitude different, such as Alaska. In states where the voter registration policy is more restrictive than the health data sharing policy, administrators might consider data release policies that favor more information.

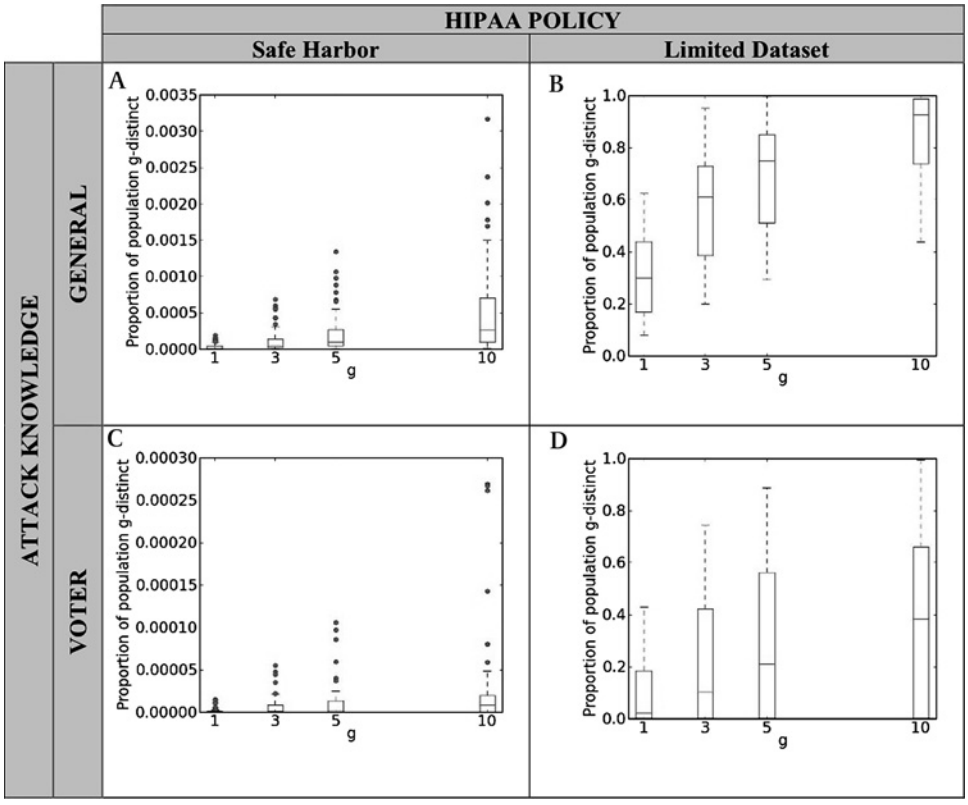

Figure 5 Distribution of *g*-distinct computations for all US states, clockwise from top left: (A) <*SAFE*, *GENERAL*>; (B) <*LIMITED*, *GENERAL*>; (C) <*SAFE*, *VOTER*>; and (D) <*LIMITED*, *VOTER*>.

The difference between the Safe Harbor and Limited Dataset risks can be seen in the trust differential, also shown in table 1. While the trust differential calculated for *GENERAL* displays a wide range, the extent of the differences is several orders of magnitude less than the differences between the trust differential for *VOTER*. For administrators using the trust differential to make data sharing decisions, this difference highlights the critical point of *VOTER* analysis for making policies that will apply across states.

Cost analysis

The estimated price per re-identification for *VOTER* is shown in table 2. The top of the table shows the states with the three minimum and maximum costs per re-identification under Limited Dataset, while the bottom shows the same for Safe Harbor. Details for all states are provided in online Appendix E. The estimated cost per re-identification under Limited Dataset ranges from \$0 to more than \$800. For the states with no charge for their voter registration lists, Virginia has the highest total risk, with an estimated 3.1 million re-identifications possible. Under Safe Harbor, the estimated cost per re-identification ranges from again, \$0, though this time with a maximum total risk of 1431 expected re-identifications in North Carolina, to a high of \$17 000 per re-identification in West Virginia. This analysis not only highlights what is possible with a particular attack, but what is likely based on these real-world constraints. Particularly for the marketer attack model, the cost and effort involved in achieving re-identifications are an important consideration.

DISCUSSION

In this paper, we introduced methods for estimating re-identification risk for various de-identification data sharing policies. We also evaluated the risk of re-identification from a known attack in the form of voter registration records. Our evaluation revealed that the differences in population distributions of US

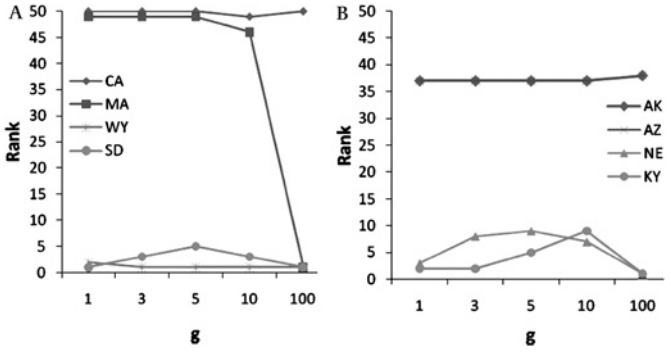

Figure 6 Ranks for top and bottom two states. (A) <*LIMITED*, *GENERAL*>; (B) <*LIMITED*, *VOTER*>.

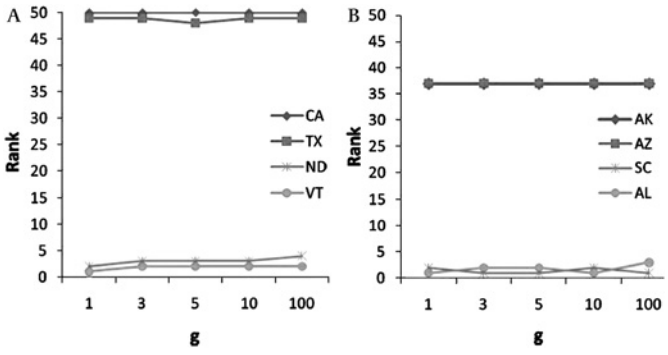

Figure 7 Ranks for top and bottom two states. (A) <*SAFE*, *GENERAL*>; (B) <*SAFE*, *VOTER*>.

Table 1 Percentage of state population vulnerable to re-identification and the trust differential according to the total risk measure

Differential rank	State	Limited Dataset	Safe Harbor	Trust differential
General				
50	DE	37.58	0.16	229
49	RI	35.25	0.13	275
48	AK	62.51	0.21	297
3	NY	25.69	0.01	3251
2	CA	19.28	<0.01	4291
1	TX	36.90	0.01	5172
Voter				
50	HI	0.01	<0.01	22
49	ND	12.38	0.01	884
48	AZ	24.61	0.02	1177
3	PA	15.31	<0.01	13088
2	VA	8.20	<0.01	12507
1	MO	36.90	0.01	5171

Table 2 Estimated cost per re-identification

State	Rank	Total risk	Price per re-id
Limited Dataset			
VA	50	3159764	US$0
NY	49	2905697	US$0
SC	48	2231973	US$0
WI	3	72	US$174
WV	2	55	US$309
NH	1	10	US$827
Safe Harbor			
NC	50	1431	US$0
SC	49	1386	US$0
NY	48	221	US$0
WI	3	2	US$6 250
NH	2	1	US$8 267
WV	1	1	US$17 000

states and their policies for disseminating voter registries lead to varying re-identification risks. Use of risk estimation approaches has the potential to improve design and implementation of data sharing policies. Here, we elaborate on some of the more pressing issues and future directions.

From theory to application

Our analysis provides a basis for comparing different privacy protection schemes both theoretically and with respect to real-world attacks. As such, the approach may be useful to privacy officials defining new policies. The difference between the *GENERAL* risk and *VOTER* risk analysis shows a wide gap between a perceived problem (the threat of re-identification using voter registration lists) and the actual results of such an attack. Furthermore, the performance of such an analysis on a state-by-state level shows that the results vary widely across the country. Data administrators in a state with a more permissive voter registration policy may wish to be more conservative in the data released, knowing the wealth of demographic information available in this single source. Comparatively, administrators in states with more restrictive voter registration policies might be interested in performing similar analyses for other available sources of identified data. They may ultimately conclude that the identified data sources that are readily available in their area are such that additional information may be included in a de-identified dataset without greatly increasing the re-identification risk. In essence, there are (at least) three different policy-making bodies that must be aware of one another: the medical data-sharing policy makers, the public records policy makers, and the data administrators making decisions about particular datasets. When making new policies or other policy-related decisions, the different policy-making bodies should be aware that their separate policies interact and their combined actions influence privacy.

Therefore, we take a moment to sketch an approach for policy makers to set appropriate protections. First, to set a specific policy, analysts should test several different policy options and document their effects on the whole population. The results of this analysis would enable the policy maker to compare policies and also to create a target identifiability range. This would define the acceptable level of risk permitted by the policy. Second, when an actual dataset is ready for release, the policy should be reexamined in light of that specific dataset. If a simple application of the policy as written leads to a risk outside the acceptable identifiability range, that dataset would be subject to further transformation before release, requiring additional suppression or retraction of certain fields. Alternatively, policy makers could authorize the release of additional fields if the estimated risk was found to be below the acceptable threshold.

Limitations and future work

The general approach of this work is limited by certain assumptions and simplifications. First, the estimates computed for the case study are only as complete as our population information. Although the US Census Bureau reports that the 2000 Census is more accurate and complete than previous censuses, the undercount rate is close to 1%.[30] Second, we used the 2000 Census as an estimate of the current population as opposed to the current population density. Third, we conflated the age reported in the Census with the year of birth reported in voter registration lists and sensitive records. For date of birth, we used a statistical model that assumes uniform distribution of birth dates. Yet, reports have shown that this may not be accurate,[33] so our estimates may misrepresent the number of distinct individuals.

Nonetheless, the idea provides several future research opportunities. First, we performed analysis for populations as a whole, but not for specific datasets. We believe a similar approach that defines the fields of intersection would be useful for dataset-specific analysis. An evaluation using a specific sensitive dataset, or multiple datasets, would allow for comparison of the theoretical risk types we evaluated here with more concrete measures. Second, this work focuses on the attack-specific risk posed by publicly available voter registration lists. While our survey provides accurate information on statewide lists, in some states voter registries are available from county governments. In Arizona, for instance, county governments are the only source for voter registration lists. Further research could show whether small counties, with more distinctive populations, or larger counties, with a lower cost per entry in the voter registries, are more vulnerable to re-identification attacks. Additionally, similar analysis could be performed with myriad other public datasets which an attacker might use for re-identification purposes.

Finally, a hurdle to the adoption of any new evaluation tool is its implementation. The risk analysis process described here can be replicated, but the implementation of such a system may be a burden. A software tool can be developed to automate the process of analyzing either a general population or a particular dataset with regard to its distinctiveness and its susceptibility to a predetermined set of attack models. We imagine that such

a tool would have information on multiple attack models, and could include different tools for estimating distinctiveness; we are in the process of developing such a tool.

CONCLUSION

This research provided a set of approaches for estimating the likelihood that de-identified information can be re-identified in the context of data sharing policies associated with the HIPAA Privacy Rule. The approaches are amenable to various levels of estimation, such that policy makers and data administrators can evaluate policies and determine the potential impact on re-identification risk. Moreover, we demonstrated that such approaches enable comparison of disparate data protection policies such that risk tradeoffs can be formally calculated. We demonstrated the effectiveness of the approach by evaluating the re-identification risks associated with real population demographics at the level of the US state. Furthermore, this work demonstrates the importance of considering not just what is possible, but also what is likely. In this regard, we considered how de-identification policies fare in the context of the well publicized "voter registration" linkage attack, and demonstrated that risk fluctuates across states as a result of differing public record sharing policies. We believe that with the methods proposed above and awareness of how different policies interact to affect privacy, a policy maker can make more informed policy decisions tailored to the needs and concerns of particular data-sets. Finally, we have outlined several routes for improvement and extension of the framework, including the incorporation of up-to-date population distribution information and application development.

Acknowledgments We thank the Steering Committee of the Electronic Medical Record & Genomics Project, particularly Ellen Clayton, Teri Manolio, Dan Masys, Dan Roden, and Jeff Streuwing for discussion and their insightful comments, from which this work greatly benefited. We also thank Aris Gkoulalas-Divanis, Grigorios Loukides, and John Paulett for reviewing an earlier version of the manuscript.

Funding This research was supported in part by grants from the Vanderbilt Stahlman Faculty Scholar program and the National Human Genome Research Institute (1U01HG00460301).

Competing interests None.

Provenance and peer review Not commissioned; externally peer reviewed.

REFERENCES

1. **Blumenthal D.** Stimulating the adoption of health information technology. *N Engl J Med* 2009;**360**:1477—9.
2. **Safran C,** Bloomrosen M, Hammond E, *et al*. Toward a national framework for the secondary use of health data: an American Medical Informatics Association white paper. *J Am Med Inform Assoc* 2007;**14**:1—9.
3. **Weiner M,** Embi P. Toward reuse of clinical data for research and quality improvement: the end of the beginning? *Ann Intern Med* 2009;**151**:359—60.
4. **National Institutes of Health.** *Final NIH statement on sharing research data* NOT-OD-03—032. February 26, 2003.
5. **National Institutes of Health.** *Policy for sharing of data obtained in NIH supported or conducted genome-wide association studies (GWAS)* NOT-OD-07—088. August 28, 2007.
6. **Sweeney L.** *Uniqueness of simple demographics in the U.S. population* Working paper LIDAP-WP4. Pittsburgh, PA: Data Privacy Lab, Carnegie Mellon University, 2000.
7. **Golle P.** Revisiting the uniqueness of simple demographics in the US population. In: *Proc 5th ACM Workshop on Privacy in Electronic Society*. 2006:77—80.
8. **U.S. Dept. of Health and Human Services.** Standards for privacy of individually identifiable health information, Final Rule. *Federal Registrar* 2002; 45 CFR, Parts 160—4.
9. **Lambert D.** Measures of disclosure risk and harm. *J Off Stat* 1993:**9**:407—26.
10. **Maynord A.** New details reveal numerous mistakes prior to election commission break-in. *Nashville City Paper* January 4, 2008.
11. **Golab A.** Social Security data puts 1.3 mil. voters at risk: suit. *Chicago Sun-Times* January 23, 2007:13.
12. **Agrawal R,** Johnson C. Securing electronic health records without impeding the flow of information. *Int J Med Inf* 2007;**76**:471—9.
13. **Fung B,** Wang K, Yu P. Anonymizing classification data for privacy preservation. *IEEE Trans Knowl Data Eng* 2007;**19**:711—25.
14. **Gionis A,** Tassa T. *k*-anonymization with minimal loss of information. *IEEE Trans Knowl Data Eng* 2009;**21**:206—19.
15. **Jiang W,** Atzori M. Secure distributed k-anonymous pattern mining data. *Proc 6th IEEE International Conference on Data Mining* 2006:319—29.
16. **Machanavajjhala A,** Gehrke J, Kifer D, *et al*. *l*-diversity: privacy beyond *k*-anonymity. *ACM Trans Knowl Discov Data* 2007;**1**:3.
17. **McGuire A,** Gibbs T. Genetics: no longer de-identified. *Science* 2006;**312**:370—1.
18. **National Research Council.** *State voter registration databases: immediate actions and future improvements, interim report*. Washington, DC: National Academy of Sciences, 2008.
19. **Samarati P.** Protecting respondents' identities in microdata release. *IEEE Trans Knowl Data Eng* 2001;**13**:1010—27.
20. **Sweeney L.** Weaving technology and policy together to maintain confidentiality. *J Law Med Ethics* 1997;**25**:98—110.
21. **Alexander K,** Mills K. *Voter privacy in the digital age* Report from the California Voter Foundation. Davis, CA: California Voter Foundation, 2004.
22. **Greenberg B,** Voshell L. Relating risk of disclosure for microdata and geographic area size. *Proc Section on Survey Research Methods, American Stat Assoc* 1990:450—55.
23. **Skinner C,** Holmes D. Estimating the re-identification risk per record in microdata. *J Off Stat* 1998;**14**:361—72.
24. **Skinner C,** Elliot M. A measure of disclosure risk for microdata. *J R Stat Soc* 2002;**64**:855—67.
25. **Truta T,** Fotouhi F, Barth-Jones D. Disclosure risk measures for microdata. *Proc. 15th International Conference on Scientific and Statistical Database Management*. 2003:15—22.
26. Princeton Survey Research Associates. *Medical privacy and confidentiality survey* 1999.
27. **Reidpath K,** Chan K. HIV discrimination: integrating the results from a six-country situational analysis in the Asia Pacific. *AIDS Care* 2005;**17**:195—204.
28. **Parker R,** Aggleton P. HIV and AIDS-related stigma and discrimination: a conceptual framework and implications for action, *Soc Sci Med* 2003;**57**:13—24.
29. **El Emam K,** Brown A, AbdelMalik P. Evaluating predictors of geographic area population size cut-offs to manage re-identification risk. *J Am Med Inform Assoc* 2009;**16**:256—66.
30. Alabama Administrative Code. http://www.alabamaadministrativecode.state.al.us/alabama.html.
31. **Mulry M.** Summary of accuracy and coverage evaluation for census 2000. *Research Report Statistics #21006—3 for U.S. Census Bureau* 2006.
32. **U.S. Census Bureau.** American FactFinder. http://factfinder.census.gov/.
33. **Schechter SE.** Toward econometric models of the security risk from remote attacks. *IEEE Security and Privacy Magazine* 2005;**3**:40—4.