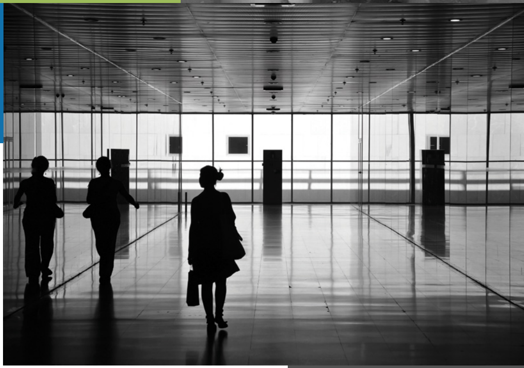
PRIVACY ANALYTICS

WHITE PAPER

Perspectives on Health Data De-identification

Learn more about health data de-identification. In this collection of articles, expert Khaled El Emam reviews various topics in the area of health data de-identification. He begins by reviewing the limits of the HIPAA Safe Harbor Standard, continues into delving in the need for big privacy when dealing with Big Data, and finally ends his discussion by explaining optimal techniques for applying masking and de-identification techniques to health datasets.



PRIVACY ANALYTICS
a QuintilesIMS company

On the Limits of the Safe Harbor De-identification Standard

FACT

De-identifying data based on Safe Harbor is often problematic, because it limits researchers' ability to perform various studies. For instance, epidemiologists often require detailed geographic information; however, Safe Harbor only permits the disclosure of three-digit zip codes, which can limit accuracy in model development and evaluation. Similarly, the field of gerontology, which is concerned with elderly individuals, is significantly hampered because Safe Harbor requires that all ages above 89 be forced into a single top coded value of 90 or greater.

Source: Beyond Safe Harbor: Automatic Discovery of Health Information De-identification Policy Alternatives http://hiplab.mc.vanderbilt.edu/people/malin/Papers/benitez_ihi.pdf

Safe Harbor has a significant disadvantage: it does not actually ensure that the risk of re-identification is low except in very limited circumstances.

Let us consider the situation where a data custodian is disclosing data publicly (for instances, online). In this case, the data custodian is creating a public use file. We are choosing this scenario because when doing a risk assessment one should consider the probability of an adversary attempting to re-identify a record. Under these conditions we do not have to worry about analyzing the probability of attempted re-identification, because with a public use file we assume that this probability is always one, which in statistical terms, means it is certain. Therefore, when creating a public file we have to assume that someone will attempt to re-identify the data because there are no controls that can realistically be put in place to prohibit that.

Going back to our example, let's say that the data custodian has applied the Safe Harbor standard to the public use file. To the right are situations where the public use file will have a high risk of re-identification even though it meets that standard.

1. The Adversary Knows Who is in the Data

Consider the simple dataset below, which has been disclosed. This dataset meets the Safe Harbor conditions in that the age is in years and only the first three digits of the zip code are given. If the adversary knows that Tom is in this dataset, and that Tom is 55, then the adversary will know with absolute certainty that the first record belongs to Tom. Here we have the highest risk of re-identification possible for this record. In fact, because all of the ages are unique just knowing the age of an individual who is in the dataset will result in re- identification with certainty.

A logical question related to this scenario is: how can an adversary know that Tom is in the dataset? There are a number of ways, as exemplified below. First, the dataset can be a population registry so everyone with a particular disease, for instance, will be in the dataset. For example, if this is a diabetes registry with everyone with diabetes in it, and if Tom has diabetes he will be in the dataset. Second, if the dataset is from a study, Tom may self-reveal by posting the information on his Facebook page that he participated in said study and is therefore in the dataset. Third, if consent to participate was required from a substitute decision maker or a parent, then that consenting individual will know that Tom is in the dataset. Finally, if the dataset is from a study of employees who volunteered and co-workers know who took the day off to participate in the study; then the co-workers would know that Tom was in the dataset and therefore, inclusion may be “knowable”.

One way to address this concern is to only disclose a simple random sample rather than a complete dataset. This adds uncertainty as to whether Tom is in the dataset or not. Of course, one has to be careful in choosing the sampling fraction, or what percentage of records to release, to ensure that this uncertainty is large enough, and that requires some analytical thought.

2. The Dataset is Not a Random Sample from the US Population

During the analysis that led to the Safe Harbor standard, the re-identification risk of datasets that meet the standard was deemed low if the dataset was a simple random sample from the US population. However, if the dataset is not a simple random sample then the risk can still be very high. Let me explain through an example of a simulation described below.

I took the hospital discharge database for the state of New York for 2007. After cleaning, removing incomplete, redundant, or duplicated information, etc., this database consists of approximately 1.5 million individuals who have been hospitalized. I then took 50% random samples of patients from that dataset and evaluated how many individual patients were unique in the sample and also unique in the population. This sampling process was repeated 1000 times and the uniqueness averaged across the iterations. Any sample meets two criteria: (a) hospitalized individuals in New York are not a simple random sample from the US population, and (b) an adversary who knows that Tom has been hospitalized would not know if he is in the selected sample or not.

If I take a cohort of males over 65 years who have been hospitalized for more than 14 days, and know only their age in years, gender, and 3 digit zip code from the sample dataset, this group had a uniqueness of 4%, and those hospitalized more than 30 days had a uniqueness of 11.14%. Therefore, by restricting or refining the cohort further and further, the individuals in the sample become more and more unique in the population of hospitalized patients in New York.

High uniqueness is bad because it means that if I find a match to Tom, it is a correct match with certainty. Even if I do not know if Tom is in the dataset, if I find a record that matches him in the disclosed Safe Harbor dataset then I will know that it is Tom’s record.

GENDER	AGE	ZIP	LAB TEST
M	55	112	Albumin, Serum
F	53	114	Creatine kinase
M	24	134	Alkaline Phosphatase

3. Other Fields Can be Used for Re-Identification

It is common for health datasets to have other kinds of fields that can be used for re-identification beyond the set included in Safe Harbor. Here we consider some examples of data elements that would pass the Safe Harbor standard but would still produce a dataset with a high probability of re-identification.

An example of a dataset that is fairly easy to re-identify is a longitudinal dataset.

Longitudinal data contains information about multiple visits or episodes of care. For example, let us consider the state inpatient database for New York for the year 2007 again, which contains information on just over 2 million visits. Some patients had multiple visits and their zip code changed from one visit to the next. If we consider that the age and gender are fixed, and allow the three digit zip code to change across visits (and the adversary knows those zip codes), then 1.8% of the patients are unique. If we assume that the adversary also knows the length of stay for each of the visits, then 20.75% of the patients are unique.

Note that length of stay is not covered by Safe Harbor, and therefore can be included in the dataset. Longitudinal information like the patient's 3-digit zip code and length of stay may be known by neighbors, co-workers, relatives, and ex-spouses, and the public for famous people. As can be seen, there is a significant increase in uniqueness when the three digit zip code is treated longitudinally, and a dramatic increase when other visit information is added to the dataset.

Although fields such as diagnosis and procedure codes are important for many analytics on health data, the reality is that this is the kind of information that an adversary would know. An adversary may not know the precise diagnosis code (e.g., ICD-9 code) of a patient, but may know the general diagnosis (e.g., the site of a cancer or that it was a heart attack). Therefore, it behooves the data custodian to consider this additional information when examining re-identification risk. Put another way, it would be difficult to justify not including this kind of information in a re-identification risk assessment. In longitudinal datasets there are many diagnoses and many procedures, which increase the risk of re-identification.

By specifying a precise and limited set of fields to consider, the Safe Harbor standard provides a simple "cookie cutter" approach to de-identification. However, it also ignores the many other data fields that can be used to re-identify individuals, reducing its effectiveness at providing meaningful universal protections for different kinds of datasets.

Conclusions

Unless the dataset being disclosed only has the fields specified in Safe Harbor, is a simple cross-sectional dataset or is a simple random sample, researchers need to be very careful about relying on the Safe Harbor standard as the basis for de-identification. It would be challenging to demonstrate that using the Safe Harbor standard ensures a low re-identification risk on many real-world datasets unless they are the most basic type of datasets.

A prudent approach, from a risk management perspective, is to follow the second HIPAA de-identification standard instead, which relies on the statistical method. This second standard can take into account the subtleties of such datasets that Safe Harbor fails to address, thus allowing data custodians to still release data but have peace of mind that they are at a low risk of re-identification.

Benefiting from Big Data while Protecting Individual Privacy

Most people would agree that we are entering the age of Big Data. This is a time where large amounts of data from multiple sources are being collected and linked together to perform sophisticated analytics for many different purposes. The data tends to be personal, in that it characterizes individual human behaviors such as their Internet surfing patterns, purchasing behavior in stores, individual health information, details on financial transactions, and physical movements, to name just a few examples. All of this personal information, especially when combined together, paints a detailed picture about individuals; their likes and dislikes, what they do, and when and where they do it

Many discussions about big data center around the technology that is needed to process such large volumes of information. Our traditional data management and data processing tools cannot handle the large volumes of data that are being collected.

Therefore, completely new systems and algorithms are being developed to process Big Data efficiently and accurately to “find the signal in the noise”. Particular challenges include extracting information from unstructured data (i.e., free form text instead of fields in a database), and linking data from multiple sources accurately to obtain detailed profiles about individuals.

> New systems and algorithms are being developed to process big data effectively and accurately to find the signal in the noise

The analytics performed on big data can be very beneficial to the individuals themselves, and to society as a whole. For example, analytics can recommend products to individuals that they may be interested in, and the recommendation might come at the time when the person may need such a product. Similarly, analytics on linked health data may identify interventions that are beneficial to people with a particular disease or condition, or detect adverse drug events that are serious enough and warrant removing a drug from the market or restricting the indications for a drug or device.

One of the questions that comes up when we talk about big data is where does all of this information come from in the first place? Some of it is customer data collected by the various organizations that are providing different products and services. Another large source of data is available freely online as individuals provide more details about their lives and interests on social networks, on blogs, and in their tweets. In some cases, it is possible to buy individual level data, for example, about magazine subscriptions or financial transactions. Government registries also provide useful information, such as date of birth information and data on things such as liens. Aggregate or summary data (e.g., averages or percentages) can be very helpful for this kind of analytics as well.

FACT

Small Business, Big Data: ”By a small but noticeable margin, executives at small companies (fewer than 1,000 employees) are nearly 10 percent more likely to view data as a strategic differentiator than their counterparts at large enterprises. Understanding the value of their data may be just what it takes to grow these small companies into large ones.”

Source: http://www.avanade.com Survey Results from “Business Impact of Big Data” by Avanade, Inc.

For example, by just knowing an individual’s zip code or postal code, it is possible to get a good estimate of an individual’s income, level of education, and number of children using just aggregate data.

Existing legal frameworks allow the collection, use, and disclosure of personal information as long as it is de-identified (or anonymized) and there is no requirement to obtain individuals’ consent if this is the case. However, the de-identification not only applies to original data, but it also applies to data that has been linked with other information. Therefore, as different data sources are integrated there is a constant need to evaluate identifiability to ensure that the risk of re-identification remains acceptably low.

One advantage of having lots of data, or big data, to analyze is that it makes de-identification easier to achieve. The reason is that there is a greater likelihood that there are more similar people in a big dataset than in a smaller one. By definition, smaller datasets are more challenging to manage from an identifiability perspective because it is easier to be unique in smaller databases.

In order to more fully understand the nuances around de-identification practice and de-identification regulations, it is important to understand the distinction between “identity disclosure” and “attribute disclosure”. Privacy laws only regulate identity disclosure which is when the identity of an individual can be determined by examining a database. For example, an “adversary” is someone who tries to re-identify a record in the dataset, can determine that record number 7 belongs to Bob Smith, then this would be considered to be “identity disclosure” because the identity of record number 7 is now known to be Bob’s.

“Attribute disclosure” is less straightforward to understand but this example, pertaining to vaccination of teen age girls against HPV (a virus that is believed to cause cervical cancer) should serve this purpose. If someone were to perform some analysis on an HPV dataset which included information on religious affiliation, they might discover that most people of religion “A” do not vaccinate their teenage daughters against HPV, because HPV is correlated with sexual activity and therefore argue that they do not need it, and then this is an example of “attribute disclosure”. Here we discovered that a particular group, characterized by their religion in this instance, has a particular attribute or behavior. Although no individual records in the database were identified, if it is known that Bob Smith follows religion “A” then no one can learn something new about him, whether he is in the database or not.

We can generalize this example to, say retail. From analyzing a large retail database linked with a magazine subscription list, we can discover that the majority of 40-year-old women, who are stay-at-home moms in zip code 12345 like tea, read a particular type of magazine and have a particular political affiliation. This conclusion does not identify any individuals, but we are still able to come to certain conclusions about these women and their lifestyles. With this information, it is possible to precisely target advertisements to these women, even though no one’s identity was revealed to draw a conclusion from the database.

As mentioned, privacy laws do not regulate attribute disclosure. Therefore, drawing inferences from databases is still a valid exercise, as long as the original data and any linked datasets are convincingly de-identified. In fact, an examination of the evidence on real world re-identification attacks reveals that they are all “identity disclosure”, which is the main type of attack that one needs to, pragmatically, protect against. But to address concerns about such inferences, transparency is important. By transparency, I mean letting the individuals know what data is being collected about them, what data is being linked or would possibly be linked to it, and how it is being used. Giving a database opt-out option would not be practical because the data would be de-identified already.

Fact: “15-20% of insurers are preparing the technology infrastructure for Big Data in the near future

Who’s Afraid of Big Data?

Much of the discussions at Big Data events were about the potential benefits of Big Data, and the really impressive ways that the fire hose of information can be used to benefit communities and create wealth.

For example, one application crawls the vast amounts of unstructured data on web sites including online news publications, social media sites like Twitter, government web sites, blogs and financial databases to predict where and when riots and protests will likely occur. They have “public” data that they access as well as commercial data. Their solution is targeted at defense and intelligence, corporate security, financial services and competitive intelligence markets. Another application monitors location and communications using mobile phones to model patterns of formal and informal interactions among employees. Yet another posts SMS numbers on products in supermarkets and stores, and when people send a message they are given a credit on their phone bill. But now the product company knows who purchased their product from their phone number and can link that with other demographic and socioeconomic data to develop a very precise profile of their customers.

Another common theme was that “people have already given up their privacy” and “the benefits are so great that privacy does not matter”. It did not seem that there is a general understanding of the privacy risks from Big Data, and that they need to be handled. While my sample is clearly not representative, from my observations these beliefs are a recurring pattern.

There are four reasons why the Big Data community needs to care about privacy. As masses of information are taken from disparate sources and mashed together to produce the desired dataset, the chances of identifying individuals and breaching their privacy becomes more and more possible. People are only now really paying attention to privacy because of the increasing amount of coverage by the media of companies like Facebook and their use of personal data. With this coverage, people are now becoming leery about the data they are sharing about themselves online. Individuals are creating accounts with fake names and dates of birth. In terms of healthcare, surveys indicate a nontrivial, and increasing, percentage of patients are lying to their doctors and omitting important details from their medical histories. Ultimately, the problem that arises with this type of practice is that the value in the data is diluted. We are no longer left with accurate information to analyze.

Organizations that are caught collecting more personal data than is necessary to provide a service or are disclosing personal information become “creepy”; and some consumers avoid continuing doing business with them, or do so reluctantly. Being known as the creepy guy in the room is not a good basis for growing a business or maintaining a consumer relationship based on analytics on their data.

By collecting personal data, organizations are also at risk of data breaches. Covered entities in the US, which are expected to follow the practices in the HIPAA Security Rule, have an annual breach rate exceeding 25% based on recent estimates. The costs of a breach of personal information are amplified when there are breach notification laws, as is the case in most states and in some Canadian jurisdictions. Breach notification costs include those from the notification itself, remedial action, litigation, lost business, and regulator penalties. These have been estimated to amount to $200 to $300 per individual affected by the breach.

FACT

U.S. Chief Technology Officer Todd Park told millions of Twitter followers that skillful use of “big data” can help patients understand their health status and provide insights into how to improve their own health. In conjunction with Big Data Week, he answered questions from around the world, focusing on practical applications for large aggregations of data, particularly in the health care field. Park, who previously held the CTO post at HHS, said privacy concerns are surmountable and that the widespread dissemination of the information can help individuals to care for themselves, as well as helping medical professionals provide better care. The key, he said, is to “make data liquid and accessible for beneficial use while rigorously protecting privacy,” which he further said is “doable.”

Source: http://hiplab.mc.vanderbilt.edu/people/malin/Papers/benitez_ihi.pdf

De-identification and Masking

Finally, regulators are weighing in more heavily on the subject of privacy. Without paying strong attention to the privacy question, stricter regulations (or legislation) will be implemented and enforced. The regulations will put limits on the collection, use, and disclosure of personal information. The extent of restrictions will, at least partially, be a function of the perceived abuses of privacy that become publicly known.

In summary, we need to tread carefully with privacy and ensure best practices are used. One of these best practices will be to de-identify data at the earliest opportunity. But this will not be the only best practice. Good governance and transparency, including some assurance against “stigmatizing analytics” will be necessary. These are the types of analytics that stigmatize individuals and affect their life opportunities, like getting a job or getting insured.

> Breach notification costs have been estimated to amount to $200 to $300 per individual affected by the breach

De-Identification and Masking

The Differences and Why it is Optimal to Utilize both Techniques to Protect Data

There has been some confusion in the health data testing expects data according to a fixed data schema, community about the difference between “masking” it is necessary to retain all the fields and have realistic and “de-identification”. This may partially be due to looking values in there. The proliferation of different terms describing the same thing, and the same terms describing different things; for example, one sees terms such “obfuscation”, “anonymization”, and “coding”. In this article I will clarify the distinction between masking and de-identification.

We need to start with a few definitions. In a dataset we make a distinction between two types of variables: direct identifiers and quasi-identifiers (also known as indirect identifiers). Direct identifiers are fields that can uniquely identify individuals, such as names, SSNs and email addresses. Direct identifiers are often not used in any data and statistical analyses that are run on the health data. Quasi-identifiers are fields that can identify individuals but are also useful for data analysis.

Examples of these include dates, demographic information (such as race and ethnicity), and socioeconomic variables. This distinction is important because the techniques used to protect the variables will depend on how they are classified.

Masking refers to a set of techniques that attempt to protect direct identifiers. There are a set of common and defensible approaches for masking direct identifiers:

1. Variable Suppression

This involves the removal of the direct identifiers from the dataset. Suppression is used more in data uses and disclosures for research and public health purposes. In those contexts it is not necessary to have the identifying variables in the dataset.

2. Randomization

Randomization keeps all of the direct identifiers in the dataset, but replaces their values with fake (random) values. If done properly, the probability of reverse engineering the masked values would be very small. The most common use case for randomization is creating datasets for software testing. This means that data is pulled from production databases, masked, and then sent to the development team for testing. Because testing expects data according to a fixed data schema, it is necessary to retain all the fields and have realistic looking values in there.

3. Shuffling

These methods take one value from a record and switch it with a value from another record. In this case all of the values in the dataset are real, but they are assigned to the wrong people.

4. Creating Pseudonyms

The creation of pseudonyms can be done in one of two ways. Both should be performed on unique patient values (e.g., SSNs or medical record numbers). One approach is to apply a one way hash to the value using a secret key (and this key must be protected). A hash is a function that converts a value to another value (the hash value) but you cannot reverse the hash value back to the original value. This approach has the advantage that it can be recreated accurately at a later point in time on a different dataset. The second approach is to create a random pseudonym that cannot be recreated. Each has utility for different use cases.

Poor Masking Techniques

Some companies employ techniques in masking tools that do not provide meaningful protection such as the following:

1. Adding Noise

The challenge with noise addition (which is most relevant for continuous variables) is problematic because there are many techniques that have been developed to remove noise out of data. Therefore, a sophisticated adversary can remove the noise from the data using various filters and recover the original values. There are many types of filters that have been developed in the signal processing domain.

2. Character Scrambling

Some masking tools will rearrange the order of the characters in a field. For example, “SMITH” may be scrambled to “TMHIS”. This is easy to reverse. To illustrate, the surname table published by the US Census Bureau in 2000 has 151,671 unique surnames. Out of the names there were 113,242 combinations of characters. There were 91,438 unique combinations of characters (i.e., they are the only name with that combination of characters). That means by just knowing the characters in a name I can figure out the name 60% of the time because the characters that make up that name are unique. As you can see this is not a reliable way to protect information.

3. Character Masking

Character masking is when the last one or more characters of a string are replaced with an asterisk. An important decision is how many characters should be replaced in such a manner. In the surname example, we replaced the last character with an asterisk. In total, there were 102,312 (~67%) of the names that still had a unique combination of characters. If two characters are replaced then 69,300 names are still unique (~46%). Without metrics to assess how many characters to replace, this type of masking may be giving a false sense of security when in fact the ability to accurately guess the name may be quite high.

4. Truncation

Truncation is a variant of character masking in that the last few characters are removed rather than replaced with an asterisk. This can also have the same risks as character masking. For example, the removal of the last character in a name still results in approximately 67% of the names being unique on the remaining characters.

5. Encoding

Encoding is when the value is replaced with another meaningless value. This process must be done with care because it is easy to perform a frequency analysis and figure out the names by how often they appear. For example, in a multi-racial dataset, the most frequent name is likely to be “SMITH”. Encoding should be performed only in the context of creating pseudonyms on unique values and not as a general masking function.

> Adding noise, character scrambling, character Masking, truncation & en-coding do not provide meaningful protection

The masking techniques that are not protective should not be used in practice. A data custodian is taking a nontrivial risk otherwise. It is important to keep in mind that even the masking techniques that are protective will reduce the utility of the data significantly. Masking should only be applied to the fields that will not be used in any data analysis, which are often the direct identifiers: fields such as names and email addresses that are not usually part of any analysis performed on the data. One should not apply masking techniques to dates or geographic information because these fields are often used in data analysis, and masking would make it very difficult to perform an analysis using those fields.

De-identification is based on characteristics of the different variables and field type. For instance different algorithms are applied to dates of birth than zip codes. A detailed discussion of the de-identification algorithms that we use can be found here - A Globally Optimal k-Anonymity Method for the De-Identification of Health Data. Because many datasets consist of both quasi-identifiers and direct identifiers, in practice it is important to apply both data masking and de-identification.

The False Promise of Data Masking

Let me start with some good news. Increasingly, I am encountering IT departments that are recognizing that they need to protect the privacy of data subjects in their databases when they use and disclose those databases for secondary purposes. Secondary purposes can be, for instance, IT sending their patient data to an outside consulting company to use for testing their business software applications. Oftentimes, IT departments will also be consulted by business lines when they have new initiatives that require the disclosure of data to external parties. It is at this time that IT should bring up the privacy issue.

However, many are still only resorting to simplistic masking techniques to achieve this privacy protection. Relying only on masking has a number of distinct disadvantages.

Masking Effectively Eliminates Analytic Utility in Data

First of all, many of the masking techniques that are commonly used will destroy the data utility in the masked fields. This means that any relationships among masked variables or between masked and non-masked variables are removed. With some masking techniques, such as shuffling, it is possible to have accurate summary statistics about a single field at a time; but not when you want to look at relationships.

For most data analytics purposes this is quite limiting.

To illustrate this, I created two fields that had a correlation of 0.8 between them. After I shuffled the two fields using the most common approach – independent shuffling, the correlation sank to zero. When I shuffled only one of them it was 1.5. Therefore, standard shuffling is not recommended if the analytics that will be performed on the data involve the investigation of relationships. But most data analysis involves the investigation of relationships.

Masking Does Not Necessarily Protect Against Identity Disclosure

Secondly, data masking methods are not necessarily protective of privacy. Protecting against identity disclosure is a legal or regulatory requirement. This means that to ensure a dataset does not contain personal information when disclosed for secondary purposes without patient consent or authorization, legal or regulatory compliance is required. For example, the HIPAA Privacy Rule states “Health information that does not identify an individual and with respect to which there is no reasonable basis to believe that the information can be used to identify an individual is not individually identifiable health information” [CFR 164.514(a)]. There exist certain expectations about how to do that. An IT department may be putting the organization at a legal or significant compliance risk position by using certain masking techniques. One cannot just make stuff up, label it as masking, and then magically it becomes acceptable to use. Let me illustrate such risks with a real example.

An organization has replaced patient identifying information in a database by creating pseudonyms, which is a data masking technique.

Unfortunately a data breach occurred and that database was lost. During the subsequent investigation the regulator working on the file concluded that despite the fact that pseudonyms were utilized, there were other demographics and diagnosis fields in the database that rendered the masking useless and showed the data to still be personal health information. This is because the risk of re-identification of the patients is quite high. Now the organization will incur the breach notification costs. Masking did not save the day.

But why did this happen? The data was masked wasn't it? Earlier in this white paper I provided some examples as to why simple masking techniques do not protect against the re-identification of patients. Let me dig deeper into this issue.

Masking techniques do not use metrics to measure what the actual risk of re-identification is, and therefore it is not always possible to know whether the transformations performed on the data were sufficient and defensible. Not using metrics can be acceptable if the masking method itself is guaranteed to ensure a low probability of re-identification. I will give you an example where it is possible to ensure that the risk of re-identification was low without explicit metrics, and one where this is not the case. In some instances we know that the probability of re-identification is going to be very small. For example, if we do random first name replacement and the database that we select from is large (say 10000 names) and the replacement names are chosen using a uniform distribution, then the probability of guessing any of the names in the database is 1/10000. This is a very small probability and the risk of reverse engineering of the randomized names will be negligible. The same can be said for techniques such as the replacement facility names and replacement addresses. Therefore, randomization is a safe data masking technique.

Simple masking techniques do not protect against the re-identification of patients

However, there will also be situations where data masking can result in data releases where the risk is high. To illustrate this, I used a common masking technique to crop the last one or two digits of the zip code. Without measuring the re-identification risk it is not possible to know whether this was protective enough or not. Let's consider an example. I used the discharge abstract dataset for the state of New York and a risk threshold of .2 (i.e. a probability equal to or less than 1.2 is acceptable). When we consider the month and year of birth, gender, and zip for all patient visits, 57.3% of the records have a probability of re-identification higher than 0.2. Cropping only one digit and retaining four digits of the zip code would mean that 25.3% of the records are high risk. If I cropped the zip code to only the first three digits, 5.5% of my records still have a re-identification risk that is higher than my threshold. By cropping without measuring the risk the data custodian would not know that more than 5% of their records have a high risk of re-identification.

Methods like cropping (which may also be called truncation) should not be used as a form of masking because you cannot know whether the data has been protected enough. Without metrics, an analyst may over- or under-truncate. The problem is that the organization may find this out at the worst possible time - when a breach has occurred.

Things to Keep in Mind

To have defensible compliance with regulations and avoid costly breaches, the general rules are:

- Only mask fields that you will not perform any analytics on.
- Since masking is not based on risk measurement, only use masking methods that can guarantee a low risk, such as random value replacement from a large database.
- For all other fields use metric-based data transformations so you can know when you have reached acceptable levels of risk that is achieved by using standard de-identification techniques.
- Both masking and measurement-based de-identification are necessary to cover all of the fields in a typical health dataset.

Otherwise the organization may be taking expensive chances with vanilla masking methods. There are many data masking techniques available today, with a key differentiator being whether they can mask static databases or can mask “on-the-fly”. In fact, neither of these criteria matter because unless the transformations done on the data, statically or dynamically, actually provide meaningful privacy protections, where and how fast you mask will not help protect the organizations from risks.

CONTACT US

251 Laurier Ave W
Suite 200
Ottawa, Ontario, Canada
K1P 5J6

Phone: 613.369.4313

www.privacy-analytics.com

sales@privacy-analytics.com