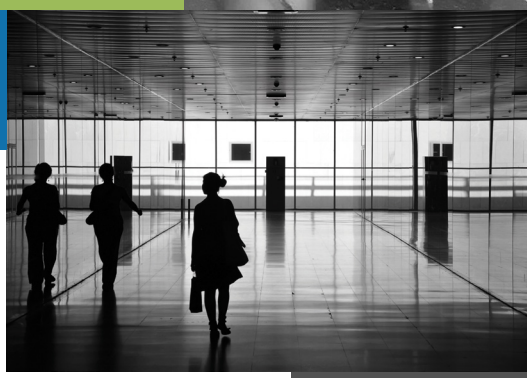
PRIVACY ANALYTICS

WHITE PAPER

Don't *Blur* the Lines between Data Masking and Real De-identification

Often confused, data masking and de-identification are not synonymous terms when it comes to unlocking protected health information (PHI) for secondary purposes. This white paper is designed to clear up the confusion by describing the true purpose of data masking with a focus on the right techniques. It also outlines the wrong techniques and limits of masking, especially when confused with data de-identification.



PRIVACY ANALYTICS

a QuintilesIMS company

Introduction

There has been some confusion in the health community about the difference between "masking" and "de-identification". This paper will clarify what masking is and offer insight into the correct techniques to use. It will also outline what techniques should not be used and the true limits of masking protected health information (PHI).

Defining the Data

In order to understand the difference between masking and de-identification, we need to understand where they apply. Masking refers to a set of techniques that attempt to remove direct identifiers in the dataset. Direct identifiers are fields that contain values that are unique to an individual and can immediately identify them, such as name, Social Security Number (SSN) or email address.

Datasets are never limited to direct identifiers; they also include quasi-identifiers (also known as indirect identifiers). Quasi-identifiers are fields that generally can't be used on their own to identify individuals but that, when linked, can allow for re-identification to occur. Direct identifiers are not often used in data and statistical analyses performed on health data. Quasi-identifier fields are also useful for data analysis. Examples of quasi-identifiers include dates, demographic information (such as race and ethnicity), and socioeconomic variables (job title, salary and education). This distinction is important because the techniques used to remove the variables will depend on how they are classified.

Masking Techniques

There are a set of common and accepted approaches for masking direct identifiers:

1. Variable Suppression

This involves the removal of the direct identifiers from the dataset. Suppression is used more for 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 that have been randomly selected from another set. 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 use in testing. Because testing expects data according to a fixed data schema, it is necessary to retain all the fields and have them contain realistic-looking values in there.

3. Shuffling

This method takes one the values from one record and switches it with a value for that same variable from another record. In this case, all of the values in the dataset are real, but they are assigned to different records.

4. Pseudonymization

Pseudonymization 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 (which must be protected). A hash is a function that converts a value to another value (the hash value). You cannot, however, 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.

Some companies employ techniques in their data masking tools that do not provide adequate defensible protection, such as the following:

1. Adding Noise

Noise addition is most relevant for continuous variables which are variables that can be measured along a continuum, like temperature, distance or height. 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 from the 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."

3. Character Masking

Character masking is the replacement of one or more characters of a string with an asterisk. An important decision is how many characters should be replaced in such a manner.

4. Truncation

Truncation is a variant of character masking in that the last few characters are removed entirely rather than being replaced with an asterisk.

5. Encoding

Encoding is replacing one value 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 in the data. For example, in a multi-racial dataset, the most frequent last 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.

Masking techniques that can easily be deciphered should not be used in practice. A data custodian may be taking undue risk with privacy otherwise. Because masking techniques are typically applied to direct identifiers, they heighten the risk of re-identification when done poorly.

It is important to keep in mind that even the protective masking techniques will significantly reduce the utility of the data. Therefore, masking should only be applied to the fields that will not be used in any data analysis. These are often the direct identifiers.

The False Promise of Data Masking

IT departments are increasingly recognizing that they need to protect the privacy of the data subjects in their databases when they use and disclose those databases for secondary purposes. Examples of secondary purposes include the following scenarios:

- Conducting comparative analysis of different healthcare institutions;
- Sharing anonymized clinical trial data with academic research groups;
- Evaluating geo-spatial information, income and physician comments using business intelligence tools;
- Examining admission and discharge dates by chronic disease and demographics;
- Monetization of claims data for resale purposes.

Many IT departments and organizations, however, are still resorting to simplistic masking techniques to try to achieve this privacy protection. Relying on masking alone has a number of distinct disadvantages.

Masking Effectively Eliminates Analytic Utility in Data

Many data 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 between fields. For most data analytics purposes this is limiting.

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. Complying with the law means that a dataset must not contain personal information when disclosed for secondary purposes without patient consent or authorization.

> Data masking methods are not necessarily protective of privacy.

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."[1] An IT department may put their organization in a position of non-compliance that risks legal action by using certain masking techniques.

Masking Does Not Use Metrics to Measure Risk

Masking techniques do not use metrics to measure the actual risk of re-identification. Therefore, it is not always possible to know whether the transformations performed on the data were considered sufficient to anonymize it and, thus, defensible. Not using metrics is only acceptable if the masking method is guaranteed to ensure a low probability of re-identification. In some instances, we know that the probability of re-identification will be very small. For example, if we do a random replacement of first names in a database that is large (say 10,000 records) and the replacement names are allocated using a uniform distribution, then the probability of guessing the correct name for any record in the database is 1/10000. This is a very small probability and the risk of reverse engineering the randomized names is negligible. The same can be said for the replacement of facility names and replacement addresses. Therefore, randomization is a safe data masking technique.

> Masking techniques do not use metrics to measure the actual risk of re-identification.

Methods like truncation should not be used as a form of masking because you cannot know whether the data has received the correct level of protection. Without metrics, an analyst may over- or under-truncate. The problem is that the organization may find this out at the worst possible time – once 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 which are not part of analytics.
- For all other fields, use risk-based data transformations so you can confirm you have reached an acceptable level of risk that is achieved by using standard de-identification techniques.
- Both masking and risk-based de-identification are necessary to cover all of the fields in a typical health dataset.

In terms of risk and the use of data, an organization is taking a potentially expensive gamble by only relying on masking. There are many data masking applications available today, with a key differentiator being whether they can mask static databases or can mask databases "on-the-fly". However, unless the transformations provide meaningful privacy protection, where and how fast you mask your data will not help protect your organization from risks.

Conclusion

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

Copyright@ 2017 Privacy Analytics

All Rights Reserved

Direct identifiers, such as names and email addresses, which are not usually included in data analysis, may be masked using a variety of techniques. Masking data completely removes all of the analytic value in indirect identifier data fields – which limit your organization's ability to gain richer analytic insight from these records.

Different methodologies exist to de-identify data but leading organizations around the world that deal with the protection of health information, including the Institute of Medicine, HITRUST Alliance, the UK Information Commissioner's Office and the Canadian Institute for Health Information, are unanimous in their endorsement of a risk-based approach to ensure proper de-identification.

Blurring the lines between masking and de-identification can be risky business. By confusing these terms, your organization is taking a big risk, one that could be costly. Masking is part of the de-identification puzzle – ensure your organization is using it correctly.

Want to learn more about risk-based de-identification? Don't miss our white paper, De-identification 101: Your Primer on Protecting Health Information.

Sources:
1. http://www.hhs.gov/ocr/privacy/hipaa/understanding/coveredentities/minimumnecessary.html