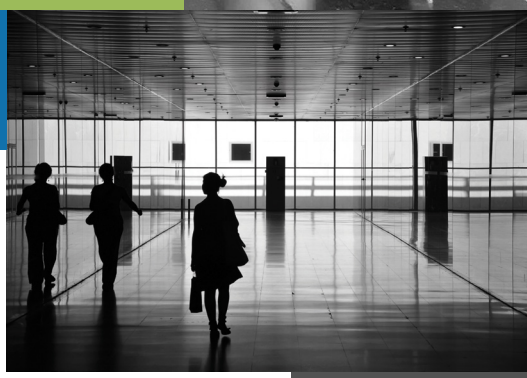
PRIVACY ANALYTICS

WHITE PAPER

Safe Harbor Vs the Statistical Method

In order to leverage protected health information (PHI) for secondary purposes, an understanding of the different de-identification mechanisms is required. Under the U.S. Health Insurance Portability and Accountability Act (HIPAA), there are two methods for de-identification: Safe Harbor and the Statistical Method (otherwise known as Expert Determination). This white paper discusses what each of these methods entail in terms of protecting your organization and how they can enable better data for analytics, research or monetization.



PRIVACY ANALYTICS

a QuintilesIMS company

Safe Harbor Vs. the Statistical Method

Two Approaches to Data Privacy

More and more, organizations are looking to leverage their data to gain valuable insights into their business and customers. In the health sector, data managers are seeking rich sources of data to better support decision-making, improve the quality of care and reduce costs. With recent technological advances, healthcare information has become increasingly available and also easier to collect, retain, use, disclose and leverage for a wide range of purposes.

The HIPAA Privacy Rule provides mechanisms for using and disclosing health data responsibly without the need for patient consent. These mechanisms center on the HIPAA de-identification standards: Safe Harbor and the Expert Determination or Statistical Method.

Safe Harbor involves data masking. It is the easier of the two methods to implement as it takes a prescriptive approach to de-identification, specifying 17 unique identifiers in the data - plus one wild card - that require masking in order for the data to be considered HIPAA-compliant. While Safe Harbor is a sound approach to preparing data for some secondary uses, more complex analytic requirements may demand the need for a risk-based de-identification methodology.

For a risk-based approach, a complete de-identification process is required. This includes masking and de-identification. An effective method for de-identification is based on HIPAA's Statistical Method. Also known as Expert Determination, this approach requires a review of the data elements by an expert who looks at the nuances contained within the data. This white paper will examine these approaches and offer solutions for how to implement the most defensible, pragmatic method for sharing health data for secondary purposes.

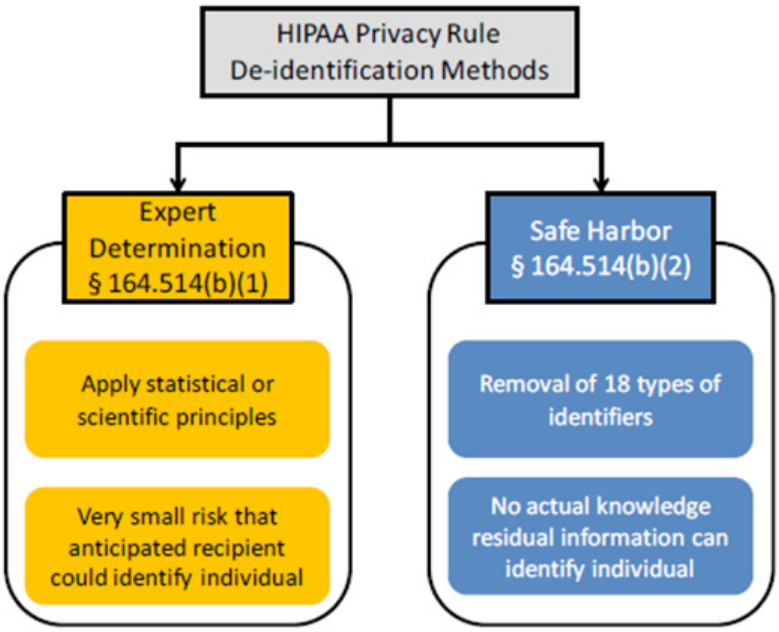


Figure 1. De-identification methods outlined by the HIPAA Privacy Rule.
Source: http://www.hhs.gov/

Safe Harbor

The Safe Harbor standard specifies 18 data elements that must be removed or generalized in a dataset. If that is done, then the dataset is considered "de-identified". Because it offers a straightforward approach, Safe Harbor is popular. Many tools are available on the market that allow organizations to quickly implement Safe Harbor at relatively low cost. Data masking is used on personal or direct identifiers in an individual's record – identifiers defined by HIPAA's Safe Harbor requirements.

Direct identifiers are fields that can uniquely identify individuals, such as name, Social Security Number (SSN) and email address. In contrast, indirect identifiers or quasi-identifiers are fields that generally cannot be used on their own to identify individuals but that, when linked, increase the risk of individual re-identification exponentially. Examples of these types of identifiers include dates, demographic information such as race and ethnicity, and socioeconomic variables like income and postal code. Quasi-identifiers are incredibly useful for data analysis. The distinction between direct and quasi-identifiers is important because the techniques used to anonymize the variables will depend on how they are classified.

Safe Harbor largely addresses direct identifiers, but direct identifiers are often not used in statistical analyses that are run on health data. Safe Harbor addresses some quasi-identifiers, but not all.

The masking techniques required by Safe Harbor do not discern or discriminate amongst those identifiers that could be used to launch a re-identification attack; rather this methodology emphasizes simplicity in order to achieve privacy. Metrics are not used 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 are, thus, defensible. Safe Harbor is useful in ensuring PHI is de-identified but is primarily applicable where analysis will be performed on basic datasets.

Datasets produced after incorporating Safe Harbor will be HIPAA-compliant, but much of the analytic utility of the data will be reduced. Thus, there are constraints with Safe Harbor. It was not conceived with longitudinal data – data collected over a period of time – in mind, allowing significant re-identification risk in these situations. Some quasi-identifiers, like occupation, are not addressed by Safe Harbor which can pose issues when unique jobs like Mayor, Governor, or even President, are present in the dataset as these individuals are easily re-identified.

A prudent approach to using and disclosing healthcare information requires de-identification of all relevant identifiers so it is important for data owners to understand the information held in their datasets and how this information will be used for secondary purposes. Where extensive and complex analysis will be performed it is important

Statistical Method or Expert Determination

that a high level of data quality is maintained with de-identification. Other, risk management-based approaches to anonymize data may be more appropriate in these situations.

Statistical Method or Expert Determination

The second standard in the HIPAA Privacy Rule is the Statistical Method, which is also referred to as Expert Determination. This standard specifies that a person, "with appropriate knowledge of and experience with generally accepted statistical and scientific principles and methods for rendering information not individually identifiable"[1], will perform the following:

1) Applying such principles and methods, determine that the risk is very small that the information could be used, alone or in combination with other reasonably available information, by an anticipated recipient to identify an individual who is a subject of the information; and,

2) Documents the methods and results of the analysis that justify such determination.

Implementing the Statistical Method is, therefore, more involved than Safe Harbor and requires specific technical knowledge about de-identification and re-identification risk.

This second standard takes into account the subtleties of the information within datasets and goes beyond the capacities of Safe Harbor in dealing with indirect identifiers. It is a robust approach in which an expert needs to consider all of the factors which would facilitate a recipient to re-identify a dataset in order to determine the level of re-identification risk. At the same time, the expert must also try to ensure that the resulting de-identified dataset will be useful for the purposes for which it has been requested.

The Statistical Method is an adaptable method of de-identification that focuses on risk management. This makes it the prudent approach in a wide variety of circumstances. The re-identification risk and de-identification methods determined for a given dataset in a particular context may not be appropriate for the same dataset in a different context or a different dataset in the same context. The determination of what is a "very small" risk is largely data and context dependent. According to the U.S. Department of Health and Human Services, the process of re-identification risk assessment and de-identification involves several steps[2]:

The Statistical method is an adaptable method of de-identification that focuses on risk management.

1) First, the re-identification risk of the data needs to be evaluated which, as noted above,

Implementing the Right Approach

is an involved process.

2) Once the risk has been measured, the expert will determine which de-identification methods should be applied to the data to minimize the risk. Depending on the needs of the data recipient and the preferences of the data custodian, appropriate de-identification methods will then be applied by the expert.

3) Lastly, the expert must measure the re-identification risk of the de-identified data to determine if the risk has been reduced to an acceptably "very small" level.

The end result of the application of the Statistical Method is robust, granular data with a minimal risk of re-identification. Complex datasets that will be used in large-scale analysis can benefit from the use of the Statistical Method de-identification standard.

Implementing the Right Approach

For organizations looking to use the Safe Harbor method, there are numerous tools available to mask data. Many of these, however, apply a blanket approach to de-identifying data that not only remove the necessary direct identifiers but that also negatively impact on the utility of date information. Under HIPAA, Safe Harbor requires all date information, except for the year, to be removed. Organizations should look for tools to help them move beyond simple masking. The most advanced Safe Harbor solutions will employ date shifting algorithms that allow date sequences and intervals to be preserved while still maintaining privacy. This enables valuable date information to be kept for use in analysis.

There are a couple of options available to an organization that wants to implement the Statistical Method. They can employ in-house statistical experts or engage de-identification consultants that are qualified to de-identify data under HIPAA. Such experts should be able to certify that the dataset has a defensibly low risk of re-identification and be able to provide an audit trail. A commercially available software tool could also be employed to conduct automated in-house de-identification.

> The end result of the application of the Statistical Method is robust, granular data.

For organizations that employ their own in-house statistical experts, there will undoubtedly be a cost associated with training these personnel and maintaining their expertise as new technologies and potential threats arise. This is in addition to the salary costs associated with maintaining an in-house expert. With regards to de-identification consultants, they may present

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

a less costly alternative depending on how often their services are required. But consults may not want to disclose their methodology to clients as this is seen as proprietary information. In this instance, an organization may not be able to prove that the methodology used to de-identify the data produced a justifiably low risk of re-identification. This could be a problem in the case of an audit, and could even put the organization at risk for a data breach.

Commercial software tools provide data custodians and privacy officers with a comprehensive and cost-effective data management solution. Automation is achievable for processes that exist at various points along the maturity spectrum of de-identification needs, from rudimentary data masking to more complete Safe Harbor implementations to the most sophisticated Statistical Method approaches. Choosing the correct solution will enable your organization to unlock the value of its personal health data.

Conclusion

Under HIPAA, there are two methods described for the de-identification of PHI. Safe Harbor, the more straightforward method to understand and implement, has constraints in terms of the quality and utility of the data that can be provided for secondary purposes. The Statistical Method allows for more robust and granular data, but it is more difficult for an organization to implement.

When de-identifying data for secondary purposes, the goals are simple: a rich and reliable source of data for analytics, research, certification or monetization. Safe Harbor provides a sound approach to de-identification for simple datasets; however, more complex data collections that contain numerous quasi-identifiers will be better served by applying the Statistical Method where a higher level of data quality can be maintained.

Sources:

1. http://www.hhs.gov/ocr/privacy/hipaa/understanding/coveredentities/De-identification/guidance.html.

2. Interpretted in Khaled El Emam and Luk Arbuckle's, **Anonymizing Health Data**, O'Reilly, 2013.