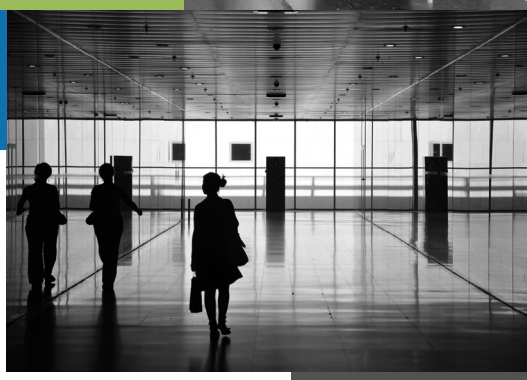
PRIVACY ANALYTICS

WHITE PAPER

How Do You Make a De-identification Expert?

The HIPAA Expert Determination method for de-identifying health data offers the most promising way to minimize privacy risk while maximizing data utility. But how can we grow the number of qualified experts with the skills to use it?



PRIVACY ANALYTICS

a QuintilesIMS company

How Do You Make a De-identification Expert?

While the HIPAA Privacy Rule only applies to the de-identification of U.S. regulated health data, it is a useful and relevant standard for the de-identification of all data. HIPAA specifies two acceptable methods for de-identifying health data. The first of these standards (45 CFR 164.514(b)(1)) is called Expert Determination. The other method, referred to as Safe Harbor, involves the removal of 18 specified identifiers,

> "A person with appropriate knowledge of and experience with generally accepted statistical and scientific principles and methods for rendering information not individually identifiable: Applying such principles and methods, determines 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..."

rather than an expert's careful analysis and determination of the risks involved in a particular disclosure. The Expert Determination method is generally considered superior because it can allow the disclosure of high-quality data while minimizing re-identification risks[1]. But who are these experts?

The need to share health information for secondary purposes in a responsible manner has become critical for many providers and businesses. It is thus imperative that there is a large pool of experts who can de-identify that data (or certify that it has already been properly de-identified).

Such experts are already in short supply. The likely downsides of a shortage are significant:

- The Safe Harbor de-identification method will be used more frequently, which will result in poorer quality data being available for secondary purposes and, sometimes, higher privacy risks.
- Many analytics on health data will not occur at all, impeding health research, public health, improvements to the health system, the growth of commercial enterprises, data-centric innovations in the delivery of care, among other benefits that would not be realized.
- Non-experts may perform de-identification improperly, resulting in the disclosure of datasets with unacceptably high re-identification risks, magnifying the privacy risks to patients and generating adverse publicity that may end up chilling research and other important secondary uses.

Without a sufficient pool of experts who can meet the growing demand for data sharing, a situation is being created that is stifling innovative data uses.

How Can That Large Pool of Experts Be Created?

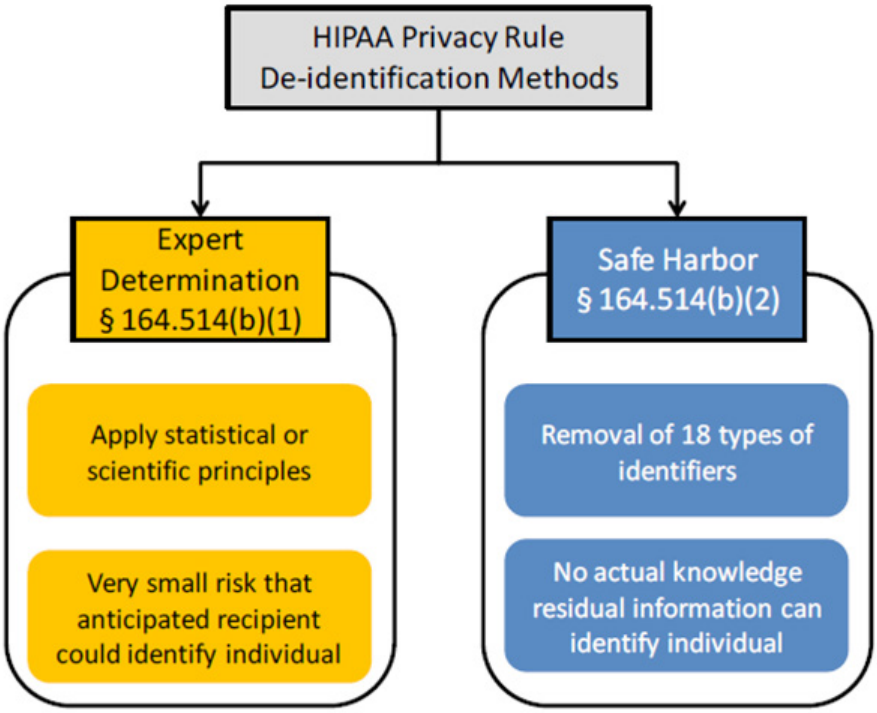


Figure 1. HIPAA Privacy Rule De-identification Methods

How can that large pool of experts be created?

The HIPAA guidance from the Department of Health and Human Services states: There is no specific professional degree or certification program for designating who is an expert at rendering health information de-identified.

Relevant expertise may be gained through various routes of education and experience. Experts may be found in the statistical, mathematical, or other scientific domains.

From an enforcement perspective, OCR [the HHS Office of Civil Rights] would review the relevant professional experience and academic or other training of the expert used by the covered entity, as well as actual experience of the expert using health information de-identification methodologies[2].

As of today, we are not aware of any university degree programs or even generally available courses that are intended to produce such experts. Furthermore, as stated in the guidance, expertise includes education and experience.

Therefore, taking a course alone would generally not be sufficient.

Given that we at Privacy Analytics perform this type of certification for a living, we are in a unique

position to propose a realistic definition of the necessary expertise. By interpreting the regulation, we can define this expertise to consist of three elements: (a) the ability to define "very small" re-identification risk in a defensible way, (b) the ability to select appropriate metrics and to measure the risk of re- identification, and (c) the ability to transform the data to ensure that the measured risk is indeed "very small." Once parts (a) and (b) have been completed, part (c) can be fully automated through software.

Once the expert determines which metric to use, measuring the actual risk in a dataset can be automated. The definition of re-identification risk metrics is well developed[3].

The expert does not need to be able to perform research on risk measurement but only to be able to measure risk using well-defined terminology and concepts.

Data transformations can also be fully automated. Transforming a dataset to reduce its re-identification risk can be done in a number of different ways (e.g. generalization by reducing the precision of variables in the data, suppression by removing some information from the data, and sub-sampling by releasing only a subset of records in the data).

Deciding on an appropriate amount of transformation (e.g. how much generalization to apply) is an optimization problem, and there are some good algorithms for doing so. De-identification experts do not need to develop these algorithms. They simply need to be able to know at a high level how they work, run them on a dataset, and interpret the outcomes. The analogy would be that data analysts and statisticians do not need to know how the numeric algorithm that performs maximum likelihood estimation works to be able to build a regression model—they just use software for that. But analysts do need to know how to specify the model and interpret the results.

To apply de- identification automation in a way that is defensible and repeatable, the expert needs to follow a methodology that:

1. Makes it possible to define what is a "very small" risk. This would involve taking into account the context of a specific data release and considering plausible methods of possible attack on the data.
2. Makes it possible to select an appropriate risk metric given the data release under consideration. For example, there are two general metrics that have been used to measure re- identification risk: maximum risk and average risk. The former is recommended for public data releases and the latter for non-public data releases.

How can an expert develop that knowledge? The solid, time-tested way of spending long years

Once the expert determines which metric to use, measuring the actual risk in a dataset can be automated

HOW DO YOU MAKE A DE-ID EXPERT?

in research and academia, learning from the literature and from other data scientists and statisticians focused on disclosure control science, and then gaining significant practical experience, generally while working with others. While this approach may be ideal, it's not very pragmatic. Current demand for de-identification experts already outstrips the supply, and this time-intensive training method will prove seriously unequal to the tasks ahead.

This is particularly true given that there are no university courses associated with this topic and therefore no existing mechanism to fill the pipeline.

As a practical matter, we have found that the following approach works well for the relatively rapid development and training of new de-identification experts:

The candidate takes a de-identification methodology course, given by a qualified expert, one which covers the most progressive standards in this field.

The course should cover the theory of how to perform the two tasks described above and include a series of practical case studies.

Subsequently, the candidate learns how to properly use de-identification software that automates the measurement and transformation. Finally, the candidate de-identifies two real datasets under the coaching and guidance of someone who is already a seasoned expert (who was involved in the de- identification of a minimum of 10 datasets).

How do we know that this proposal is a reasonable one? Because we have found that individuals who go through this process produce outcomes that are the same or very similar to the outcomes that the seasoned experts produce.

Since de-identification is our business, we have many of these seasoned de-identification experts working with us. We know what the expected outcome that an expert would produce looks like, and we have learned that experts developed using this approach produce equivalent outcomes.

The background of the candidate can be data analysis, database management, health data management, statistician, or software programming. Non-expert observers are frequently surprised at how objective the overall de-identification process is. Notwithstanding, there is some degree of subjectivity. That subjectivity can be reduced by applying well-defined methods to reach consensus among those who understand the data and its risks in particular settings. For example, there may be

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

some discussion or uncertainty about whether a certain data field can be used to re-identify a data subject.

There are well-established ways to obtain a reasonable consensus opinion to resolve that question and allow the de-identification to proceed, for example, by organizing consensus workshops.

Privacy Analytics supports that process by providing office hours for our clients where our experts can provide feedback and input as well.

This process can help responsible data holders develop the internal expertise to de-identify datasets and ensure that the risk of the data being re-identified is “very small.” Following an established method of developing de-identification experts will help protect privacy while unleashing the analytic power of data.

Sources

1. K. El Emam, Risky Business: Sharing Health Data While Protecting Privacy. Trafford, 2013.
2. Guidance Regarding Methods for De-identification of Protected Health Information in Accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule. U.S. Health and Human Services Office of Civil Rights, Nov. 26, 2012, available at http://www.hhs.gov/ocr/privacy/hipaa/understanding/coveredentities/De-identification/hhs_deid_guidance.pdf.
3. K. El Emam, Guide to the de-identification of personal health information. CRC Press (Auerbach), 2013.