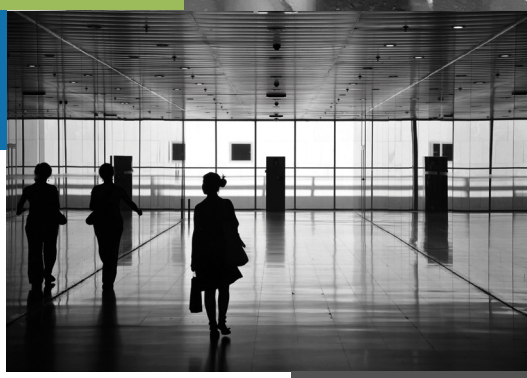
PRIVACY ANALYTICS

WHITE PAPER

De-identification 401

Critics who question the value of de-identification often overlook the quality of the de-identification process. The issue is not whether there is value in de-identification but whether de-identification has been done properly so that data is truly anonymous. Not all de-identification practices are created equal. An approach that uses the Expert Determination method, and that requires the expertise of individuals with knowledge of the statistical and scientific principles of de-identification, is the optimal approach.



PRIVACY ANALYTICS

a QuintilesIMS company

An Optimal Approach to Data De-identification

Expert Determination focuses on the risk that an individual could be identified in the data. In order for personal information to be considered protected, the chances of an individual being re-identified from the data must be "very small". How we define a very small risk will change depending on the content of the data and the context of its use. Determining a dataset's exposure to risk is done case by case and requires an examination of factors like privacy and security practices at the data recipient's site, the sensitivity of the information contained in the data and who will access it.

The goal of a quality process is to balance the need for privacy against the need for precise data. This is accomplished by selecting the identifiers that are most important to the analysis so that greater specificity is retained for these values. As assessment of the various types of re-identification attacks will then govern the extent of de-identification to the data. The involvement of experts ensures a rigorous process that complies with legislative requirements.

Is De-identification Worthwhile?

Critics of de-identification say that there is little point in trying to make personal information anonymous. They will cite academic studies which claim that re-identifying an individual from de-identified data is a relatively simple process[1]. The conclusions drawn from these studies make assumptions and broad generalizations, however, which are not supported by the research[2].

Unfortunately, this can leave the impression that de-identification is not a worthwhile practice. The question is not whether there is value in de-identifying personal information; it is whether or not data that is said to be anonymous is, in fact, truly anonymous. As discussed previously in this White Paper series, the removal of direct identifiers like name and address from a dataset is not enough to ensure that the data is actually de-identified. Effective de-identification requires a comprehensive approach and careful understanding of the data that comes from having expertise in the science and methodology of de-identification.

Even though it is impossible to guarantee that re-identification of an individual could never occur, data that is de-identified using a process based on Expert Determination faces a minimal risk of re-identification. Many highly regarded organizations, including the Institute of Medicine[3], HITRUST[4], the UK Information Commissioner's Office[5] and the Canadian Institute for Health Information[6], have all identified a risk-based approach to de-identification, like Expert Determination, as the optimal approach.

This is the fourth and final paper in our series exploring de-identification and the techniques used to protect patient privacy. This paper describes the steps involved in an effective de-identification process. It examines the factors influencing re-identification risk, how to determine

What Impacts the Risk of Re-identification?

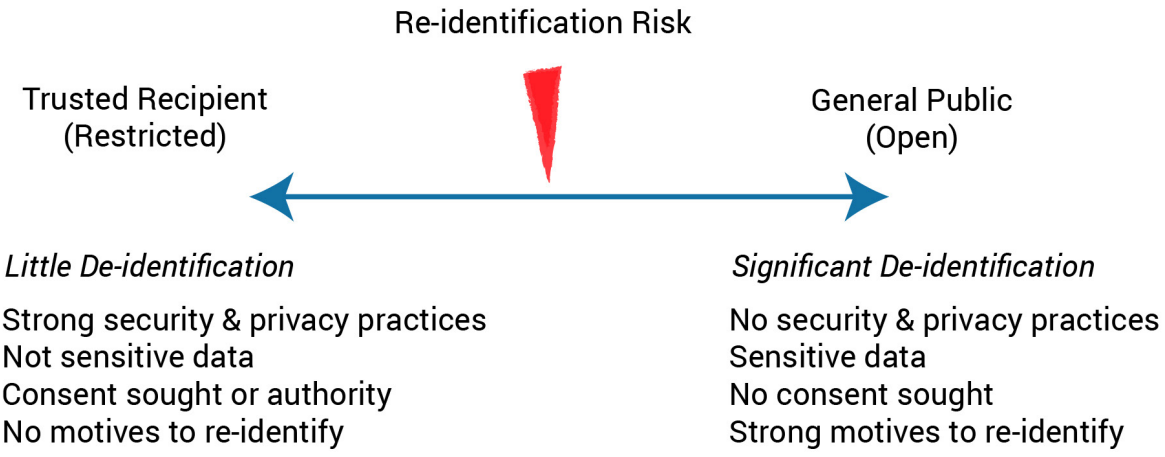


Figure 1: The spectrum of de-identification

the data's risk exposure and walks through a proven process for real data de-identification. It is not possible to have a zero risk of re-identification. Thus, the aim in de-identifying data is to manage the risk so it is as small as possible under the given circumstances. The threat of re-identification posed in any situation is influenced by a variety of factors. In this section we will explore the factors that impact this risk. HIPAA legislation states that de-identification should achieve a "very small" risk of re-identification. The ability to re-identify an individual in a dataset sits somewhere along a spectrum that ranges from an almost zero chance of re-identification at one end to almost certain re-identification at the other. Moving along the spectrum towards a smaller probability of re-identification means that less de-identification is necessary. Moving towards a higher probability for re-identification means that significant de-identification is needed. The goal is to find the correct balance in any situation.

Finding the correct balance is influenced by who will use the data (a trusted recipient versus the general public), the security and privacy practices in place at the data recipient's site, the level of sensitivity of the data and the motivations to re-identify the data.

If a dataset will be used by an organization with strong security and privacy practices and it is not highly sensitive data, the risk of re-identification is considered to be small. Protecting the privacy of individuals in this situation can be achieved, in part, through contracts and security protocols since we are dealing with trusted recipients. Thus, less data manipulation is needed to de-identify individuals. However, if the dataset will be disclosed to a third party that has little or no privacy and security practices and has a strong motive to re-identify highly sensitive data, contractual protections are not practical. In this situation, significant de-identification of the data is necessary to protect individual privacy.

Determining the Exposure to Risk

Figure 2: Risk Thresholds

As a result, the correct "amount" of de-identification changes with respect to the conditions and the context of the data's use.

This next section explores how much risk we are prepared to face when sharing data. In the last section, we learned that the risk of re-identification is impacted by the context in which the data will be used – who will be using it, how it will be protected and what it contains. Here we will further examine the factors that influence risk so that we may figure out the data's exposure to re-identification.

Figure 2 shows that there are three factors that concern us in assessing the level of risk. Mitigating Controls look at the privacy and security practices of the data recipient, Motives & Capacity gauge the skill and desire that the recipient has to re- identify the data, and Invasion of Privacy judges the sensitivity of the data and the potential for harm should the data be breached. To determine what the data's exposure to risk is in a given situation, we follow a stepwise approach that will lead us to choose an appropriate value based on the context of the data release.

Step 1: Mitigating Controls

The first step evaluates the mitigating controls that a data recipient has in place. This looks at the security and privacy practices used by the data recipient to ensure the data's protection from unauthorized access. The higher the mitigating controls, the better the security protocols that are in place and the lower the re-identification risk. This dimension is scored as public, low, medium or high.

Step 2: Motives & Capacity

Next, we assess the motives and capacity of the data recipient to re-identify the data. Motives include stealing data for financial gain, curiosity, and the desire to show that re-identification is possible. Capacity looks at whether the data recipient has the skills and know-how to successfully re-identify individuals in the data. This dimension is scored as low, medium or high.

Step 3: Invasion of Privacy

The next dimension evaluates the Invasion of Privacy and is characterized by the extent to which a disclosure would be an invasion of privacy for the individual involved. Here there are three considerations: a) the sensitivity of the data (the greater the sensitivity, the greater the invasion of privacy) b) inappropriate processing of the data leading to potential injury (the greater the potential for injury, the greater the invasion of privacy) and c) whether the appropriate consent to disclose the data was received. The Invasion of Privacy is lower when consent has been obtained. This dimension is scored as low, medium or high.

Step 4: Risk Exposure

The final step is to determine the exposure to risk based on strong precedents from reputable institutions that release data for secondary purposes, such as a Department of Health or the Centers for Disease Control. The resulting scores from the Mitigating Controls, Motives & Capacity and the Invasion of Privacy dimensions (low, medium or high) are then mapped to these precedents to determine the appropriate measure of risk for the given situation.

This last section brings together the knowledge acquired over this White Paper series. The de-identification process outlined here uses the risk-based approach of Expert Determination. The involvement of experts in this process ensures that it is rigorous and complies with legislative requirements.

An Optimal Approach to De-identifying the Data

Step 1: Selection and Ranking of Quasi-identifiers

To begin the process, the quasi-identifiers that are to be released in the dataset are selected and ranked in their order of importance[8]. This is the order of importance for the person conducting the research or analysis on the de-identified dataset. By ranking the quasi-identifiers, the data's usefulness can be maximized while balancing it against the re-identification risk so that optimal anonymization is achieved. Let's take, as an example, the two quasi-identifiers of income and zip code. If income is ranked as the most important quasi-identifier and zip code as the least important one, then the transformations performed on the data will try to make the least change to the variable income so that the specificity of the values remains high. Greater changes will be permitted to the variable zip code

since its importance to the analysis is low. As a result, a significant amount of specificity on zip code may be lost.

Step 2: Determining the Risk Exposure

Measuring the risk contained in the raw data is an essential part of the Expert Determination method. This process was outlined in the previous section. By finding the data's Risk Exposure we are able to ascertain the overall level of re-identification risk.

Step 3: Measuring the Risks to the Data

Once the Risk Exposure is identified, a risk analysis of the dataset is performed. This may involve the services of a de-identification expert who will perform a detailed risk assessment under various threat models and scenarios such as re-identification attacks based on prosecutor risk, journalist risk and marketer risk[9]. Various scenarios are assessed to calculate their potential to re-identify an individual from the data. De-identification techniques will then be applied to the data to a greater or lesser degree in order to address the Risk Exposure with respect to the various types of attack.

Step 4: De-identifying the Data

To reduce the Risk Exposure, de-identification techniques are applied to the variables in the dataset. This will mainly involve aggregating data values, particularly for the variables ranked as less important, and suppressing cells or records in the dataset[10]. The degree of de-identification needed to achieve an acceptable level of risk will determine how many ranges are used when aggregating values and the extent of cell and record suppression. If more de-identification is needed, for example, values will be aggregated more so that there are fewer groups for a given variable. For example date of birth could be aggregated to the month and year of birth (e.g. February 1960) or further de-identified so that only the year of birth is provided (e.g. 1960). When data is properly de-identified, society benefits from the insights that come from sharing data with analysts and researchers while still ensuring that the privacy of the individuals in the data remains intact.

Conclusion

De-identification is a vital tool in the protection of privacy. However, the mechanics of de-identification are complex. In many instances, data that is reputed to be de-identified is not truly anonymous which paves the way for re-identification of individuals and privacy breaches.

Privacy legislation in the U.S. and other jurisdictions set out specifications to limit the risks to personal privacy for data that is used or disclosed for secondary purposes. 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, the UK Information Commissioner's Office and the Canadian Institute for Health Information, are unanimous in their endorsement of a risk-based approach like Expert Determination to ensure proper de-

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

identification.

Sharing data with researchers and analysts for secondary uses benefits society as a whole. The answer to re-identification risk is not to limit the sharing of data; rather, it is to apply the science and use methods that ensure data is reliably de-identified. This requires not only commercial software tools to deliver an automated and repeatable process but also experts in de-identification who can perform a certified assessment of re-identification risks, evaluate the strengths and weaknesses of your organization's current processes for legislative compliance and help your company establish best practices in protecting personal information.

Interested in learning the benefits of incorporating risk-based de-identification? Look to ASCO CancerLinQ's example. Using Privacy Analytics' software, they have built the world's leading Learning Healthcare System. Read the case study here: https://www.privacy-analytics.com/files/Asco-Case-Study.pdf.

Appendix: Terminology

TERM	DEFINITION
Aggregation	Interchangeable with the term Generalization. Involves grouping values within a data field so that a less precise, but still accurate, value is assigned. For example, a birth date of May 10, 1956 can be assigned the aggregate birth date value of May 1956 or birth dates could be even further aggregated so the value assigned is 1956.
Cell Suppression	Removing a value from a single field (cell) of a record in the dataset when its inclusion presents a high risk of re-identification, e.g. a field in a patient record that specifies a very rare disease could be suppressed.
Dataset	A collection of related data records. Most commonly, a dataset refers to the contents of a database with many tables of data, where every column in the table represents a particular variable.
De-identification	A process that removes or suppresses, and/or alters personally identifiable information in a data collection so that it may be shared within the organization, with other organizations, or individuals for secondary purposes. This term is sometimes used interchangeable with the term anonymization
Direct Identifier	The fields within a dataset that can easily be used alone to uniquely identify individuals. This includes information such as name or email address.
Expert Determination	Also referred to as Statistical Method. A standard methodology for de-identification specified under the HIPAA Privacy Rule. Expert Determination requires a person with appropriate knowledge of, and experience with, generally accepted statistical and scientific principles and methods to certify and document that a dataset is sufficiently de-identified such that there is a very small risk that an individual can be identified from the data.
HIPAA	The Health Insurance Portability and Accountability Act. Federal legislation enacted in the United States in 1996 that protects the confidentiality and security of personal healthcare information by setting limits on the use and disclosure of a person's data unless consent for additional secondary purposes has been obtained from the individual subject of the information (or is compelled by court order).
Journalist Risk	The risk of re-identification posed by an attacker attempting to re-identify a single individual in the data using information they know about the individual, usually a family member, friend, colleague or a well-known person. The attacker does not know with certainty that the individual they are trying to identify is in the dataset.
Marketer Risk	The risk of re-identification posed by an attacker who attempts to re-identify as many individuals as possible in a dataset. The attacker is not concerned if some of the individuals are incorrectly re-identified only that as many of the individuals as possible are re-identified.
Protected Health Information (PHI)	Health data that can be used to uniquely identify or locate an individual. Examples of protected health information include health plan numbers, disease diagnoses, hospital admissions information or lab results.
Privacy Breach	The result of unauthorized access to, or collection, use and disclosure of personally identifiable information for one or more individuals whose information is contained in a dataset. Privacy breach may be inadvertent (security vulnerability) or intentional (hacker).

TERM	DEFINITION
Prosecutor Risk	The risk of re-identification posed by an attacker attempting to re-identify a single individual in the data using information they know about the individual, usually a family member, friend, colleague or a well-known person. The attacker knows with certainty that the individual they are trying to identify is in the dataset.
Quasi-identifier	Fields within a dataset that can be used in combination with one another to identify individuals. For example, birth date or postal code. Quasi-identifiers are also referred to as indirect identifiers.
Re-identification	The identification of a unique individual within a dataset that was supposed to have been de-identified.
Secondary Use	Any use of a dataset other than for the provision of direct patient care. Secondary uses of healthcare data include research, analysis, quality and safety, accreditation, policy setting, and marketing or other business applications.

Sources:

1 De Montjoye, Yves-Alexandre et al. (2015). Unique in the shopping mall: On the reidentifiability of credit card metadata. Science, 347(6221), 536-539. doi: 10.1126/science.1256297.

2 El Emam, Khaled. (2015, February 6). Is it safe to anonymize data? The BMJ. Retrieved from http://blogs.bmj.com/bmj/2015/02/06/khaled-el-emam-is-it-safe-to-anonymize-data/

3 Institute of Medicine of the National Academies. (2015). Sharing Clinical Trial Data: Maximizing Benefits, Minimizing Risk. Retrieved from the National Academies Press. http://www.nap.edu/catalog/18998/sharing-clinical-trial-data-maximizing-benefits-minimizing-risk

4 https://hitrustalliance.net/hitrust-csf/?

5 Information Commissioner's Office. (2012, November). Anonymisation: managing data protection risk code of practice. Retrieved from the UK Information Commissioner's Office. https://ico.org.uk/media/1061/anonymisation-code.pdf

6 Health System Use Technical Advisory Committee. (2010, October). "Best Practice" Guidelines for Managing the Disclosure of De-Identified Health Information. Retrieved from the Electronic Health Information Laboratory. http://www.ehealthinformation.ca/wp-content/uploads/2014/08/2011-Best-Practice-Guidelines-for-Managing-the-Disclosure-of-De-identificatied-Health-Info.pdf

7 The motivations to re-identify data were explored in depth in the third paper of this series, De-identification 301: Three Adversaries Who Could Attack Your Data.

8 A complete discussion of quasi-identifiers and direct identifiers can be found in De-identification 201: Fundamentals of Data De-identification.

9 For an explanation of the various types of re-identification attacks, see De-identification 301: Three Adversaries Who Could Attack Your Data in this series.

10 A discussion of de-identification techniques can be found in De-identification 201: Fundamentals of Data De-identification.