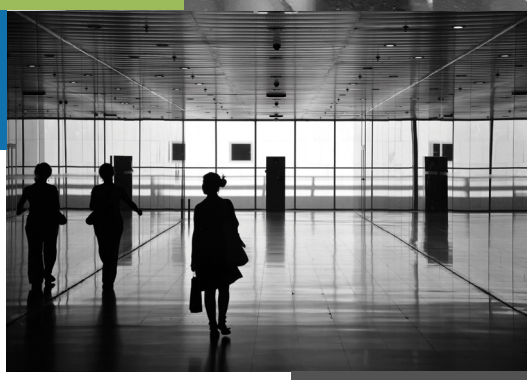
PRIVACY ANALYTICS

WHITE PAPER

De-identification 201

Only a handful of experts exist around the world who are qualified to manually de-identify data. This is because de-identification is a complex and challenging field that requires highly specific knowledge. Simply removing the names and other types of direct identifiers from a dataset is insufficient to achieve de-identification. The data will also contain other indirect or quasi-identifiers, such as age, date of birth and zip code that, when combined, can be used to positively identify an individual.



PRIVACY ANALYTICS

a QuintilesIMS company

Fundamentals of Data De-identification

There is a trade-off between maximizing privacy and maximizing the usefulness of the data for secondary purposes. Balancing these two things is achieved by applying a combination of data masking and data de-identification techniques.

While masking largely removes direct identifier values, de-identification helps to preserve the values of quasi-identifiers by using approaches like generalization and sub-sampling. Unique records, or outliers, are suppressed or removed from the data.

The HIPAA Privacy Rule, part of the U.S.'s HIPAA legislation covering the use and disclose of protected health information, provides two standards for de-identification: Safe Harbor and Expert Determination. Safe Harbor is an easy-to-follow approach for de-identification but has significant drawbacks since extensive information can be lost in using this method. Expert Determination, on the other hand, requires the knowledge of experts to assess the risk of re-identification and is based on the application of current research in this area. This method helps to preserve the analytical quality of the data.

The Challenge of De-identification

The goal of de-identification is straightforward – to ensure that data used beyond its primary intent cannot be matched to the person it describes so that their privacy is protected. The execution, however, can prove to be complex and challenging. Many people assume that simply removing names, addresses and other identifiers (like Social Security Number) should be sufficient to make information anonymous. Yet, data will contain other personal details that, while not obviously identifying, can be used to re-identify a person. This includes information like date of birth, marital status, occupation and even movie choices.

A few years ago, Netflix launched a competition inviting developers to improve on its algorithm that provides movie recommendations. For the inaugural Netflix Prize, the company released 100 million supposedly de-identified records that showed customers' ratings for the movies they'd watched. To release this data publicly, Netflix replaced customers' names with pseudonyms. It took two researchers little more than two weeks to re-identify certain targets in the data by matching these ratings to the movie ratings that those individuals had provided to another public movie database, IMDb. This privacy breach resulted not only in a public relations nightmare for Netflix but a lawsuit which was later settled and the cancellation of a similar contest the following year[1].

When data is not de-identified in a comprehensive way, re-identification becomes possible. It is important that the approaches used to de-identify data are effective in making the risk of re-identification very small and that these methods can be defended under scrutiny. However, there are only a handful of experts in the world who are qualified to manually de-identify data. As a result, research into automated, risk-based approaches to de-identification remains active.

This paper, the second in a series that explores de-identification, looks at the different types of

Assessing Data for De-identification

identifiers, discusses and de-identification and masking and how they work together, examines the HIPAA standards of Safe Harbor and Expert Determination for ensuring health data privacy.

Assessing Data for De-identification

In assessing data for secondary uses, data custodians need to know the content of their datasets and understand the context in which the data will be shared.

There is a necessary trade-off between privacy protection and data quality when sharing information for secondary uses. To determine the correct balance between these two, it is imperative to know the context in which data will be shared and how it will be used. As Figure 1 illustrates, it is impossible to provide both maximum privacy and maximum usefulness. The only way to guarantee optimal privacy (max privacy) is to remove all values from the dataset, rendering it useless (no utility). Understanding how data will be accessed – will it be shared with a trusted researcher who has signed a confidentiality agreement or will it be posted to a public website – helps to determine where along the acceptability curve to aim your de-identification efforts. If in doubt, it is preferable to err on the side of privacy protection which means greater de-identification.

Direct Identifiers versus Quasi-identifiers

Beyond knowing the context in which data will be shared, data owners must know the identifiers in their data and the degree to which they can be used to uncover an individual's identity. Figure 2 shows that identifiers are classed as either direct identifiers or quasi-identifiers (indirect identifiers).

Whether or not a piece of information is an identifier is based on three characteristics. An identifier must be replicable, distinguishable and knowable.

If a field in a dataset has values that are stable over time, it is considered replicable. A health plan number is a good identifier because its value is unlikely to change over time, while something like blood sugar level that goes up and down over the course of a day or a month is not a good identifier. An identifier being distinguishable refers

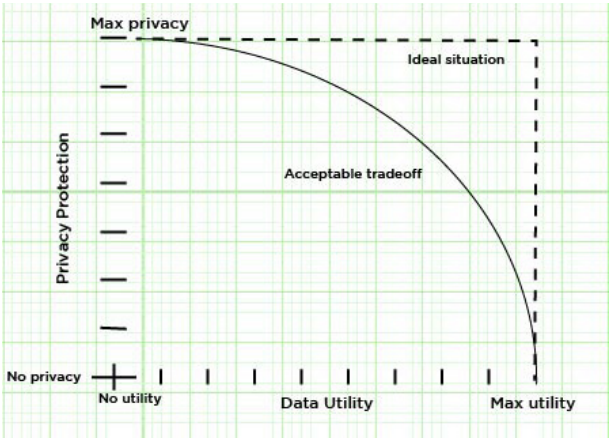


Figure 1: The trade-off between privacy protection and data utility.

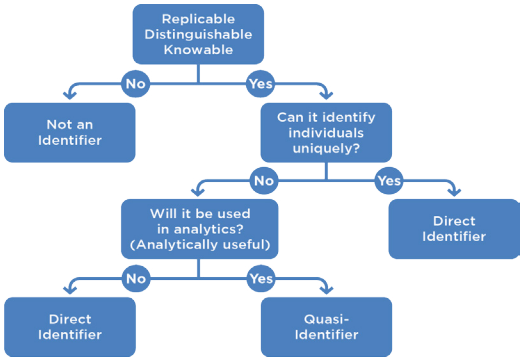


Figure 2: Decision Tree to determine type of identifier (direct or quasi)

Data Masking	De-identification
✓ Reduces risk of re-identification	✓ Reduces risk of re-identification
✓ Data transformation technique	✓ Data transformation technique
X No concern for analytics	✓ Data retains a very high analytic value

Figure 3: Characteristics of data masking and de-identification

to the fact that there is sufficient variation in the values across the dataset to distinguish among individuals. A diagnosis of diabetes would not be distinguishable in a dataset of diabetes patients, although it may be in a dataset of a general population. Knowable means that the identifier must describe information that can be known and then used to re-identify the records in a dataset. Someone wishing to re-identify a record could know this information because they are acquainted with individuals in the dataset or because the information exists publicly (e.g. a voter registration list).

Direct identifiers are data fields that can be used alone to uniquely identify individuals. This includes elements such as name, email address or Social Security Number, where each of these is generally associated with only one person. Census results are de-identified and provided/sold to third parties for further analysis. Open data initiatives are focused on unleashing the potential of the data for the creation of innovative products and services, for creating transparency, to increase service offerings to citizens or to allow citizens to have more control over their healthcare.

Quasi-identifiers are fields in a dataset that can be used in combination with one another to identify individuals. Examples include gender, zip code, birth date, profession and income. While there are many people who share the same gender, birth date or ZIP code, the combination of these for any one person may be unique, particularly if that person resides in a rural area with a small population. In fact, a study conducted at Carnegie Mellon University found that the majority of the U.S. population could be uniquely identified based on their gender, age and zip code[2].

The distinction between direct and quasi-identifiers is an important one as this will determine which elements are anonymized through masking and which ones are anonymized through de-identification.

Masking and De-identification

Although data masking and de-identification are often grouped together for discussion, the two use different approaches to making data anonymous. As we just learned, masking and de-identification deal with different identifiers in the dataset. Masking is used to anonymize direct identifiers while de-identification is used to anonymize quasi-identifiers.

While both of these approaches use various techniques to change the data, such as suppression or the use of pseudonyms, only de-identification is concerned with retaining data's analytic value for secondary uses.

In the Netflix example, mentioned earlier in this paper, masking was used to remove customer names from the database. However, as that case showed, masking is often not enough to

Masking and De-identification

De-identification Techniques

Record Suppression: Removing a record from the dataset when the combination of quasi-identifiers presents a high risk of re-identification.

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 record that specifies a rare disease.

Sub-Sampling: Taking a random sample of records from a dataset. For example, if the requirement is for a dataset of 1000 records, this could be achieved by taking a random sample of 10% of the records from a larger 10,000 record dataset.

Aggregation or Generalization: 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 1956.

effectively prevent re-identification.

Furthermore, because masking tends to rely on techniques that get rid of data, it can distort the information and reduce the data's usefulness.

The aim of de-identification is to do as little as possible to alter the data while still effectively making the information anonymous. De-identification uses techniques like record suppression, cell suppression, sub-sampling and aggregation to transform the data values while minimally distorting the data.

Using Masking and De-identification Together

In practice, masking and de-identification are used together to optimize the balance between protecting privacy and maintaining the usefulness of the data. Direct identifiers in the data are masked while quasi-identifiers are de-identified.

To illustrate this, below is an original dataset prior to anonymization that contains the direct identifier Name and two quasi-identifiers: Gender and Year of Birth. The dataset also contains a column showing

Original Database				
	Direct Identifier	Quasi Identifiers		
ID	Name	Gender	Year of Birth	Gene Mutation
1	Bruce Wayne	M	1939	-ve
2	Diana Prince	F	1941	-ve
3	Barbara Gordon	F	1961	-ve
4	Clark Kent	M	1938	-ve
5	Peter Parker	M	1962	+ve
6	Bruce Banner	M	1962	+ve
7	Natasha Romanoff	F	1964	-ve
8	Tony Stark	M	1963	-ve
9	James Howlett	M	1974	+ve
10	Clint Barton	M	1964	-ve
11	Sue Storm	F	1961	+ve
12	Reed Richards	M	1961	+ve

Table 1: Original database containing direct and quasi-identifiers.

Privacy Standards: Safe Harbor versus Expert Determination

sensitive information about the individual that they may not want to have disclosed – the results of a test for a genetic mutation.

In order to make this data anonymous, the direct identifier Name is removed from the data and the quasi-identifier Year of Birth is generalized to the level of decade. Two records have also been suppressed in this dataset – those of Diana Prince and James Howlett (see Table 2). This is because the combination of their quasi-identifiers (a female born in the 1940's and a male born in the 1970's) made them unique cases, or outliers, in this dataset and thus easy to re-identify.

The HIPAA Privacy Rule

Many jurisdictions around the world have enacted legislation to Protect Personally Identifiable information (PII) and Protected Health Information (PHI). Under the U.S.'s Health Information Portability and Accountability Act (HIPAA), the HIPAA Privacy Rule sets out the standards for the use and disclosure of PHI held by Covered Entities and their Business Associates. Covered Entities are health plans, healthcare providers and data clearinghouses, while Business Associates are people or organizations that do work on behalf of a Covered Entity and require the use or disclosure of PHI. Examples of Business Associates include third-party administrators and claims processors for health plans, attorneys that have access to their clients' PHI, or third-party researchers.

The HIPAA Privacy Rule provides mechanisms for using and disclosing health data responsibly without the need for patient authorization. These

De-identified Database Quasi-identifiers			
ID	Gender	Year of Birth	Gene Mutation
1	M	1930-1939	-ve
2	F	1960-1969	-ve
3	M	1930-1939	-ve
4	M	1960-1969	+ve
5	M	1960-1969	+ve
6	F	1960-1969	-ve
7	M	1960-1969	-ve
8	M	1960-1969	-ve
9	F	1960-1969	+ve
10	M	1960-1969	+ve

Table 2: De-identified version of database from Table 1

mechanisms center on the Rule's two de-identification standards: Safe Harbor and the Expert Determination or Statistical Method.

An Overview of Safe Harbor

The Safe Harbor method uses a list approach to de-identification and has two requirements[3]:

1) The removal or generalization of 18 elements from the data. (See Table 3)

2) That the Covered Entity or Business Associate does not have actual knowledge that the residual information in the data could be used alone, or in combination with other information, to identify an individual.

Safe Harbor is a highly prescriptive approach to

Safe Harbor Direct and Quasi-Identifiers

1. Names
2. Zip Codes (except first three character)
3. All elements of dates (except year)
4. Telephone Numbers
5. Fax Numbers
6. Electronic Mail Addresses
7. Social Security Numbers
8. Medical Record Numbers
9. Health Plan Beneficiary Numbers
10. Account Numbers
11. Certificate/License Numbers
12. Vehicle Identifiers and Serial Numbers (including license plate numbers)
13. Device Identifiers and Serial Numbers
14. Web Universal Resource Locators (URL)
15. Internet Protocol (IP) Address Numbers
16. Biometric Identifiers, including finger and voice prints
17. Full-face photographic images and any comparable images
18. Any other unique identifying number, characteristic or code

Table 3: The 18 different identifiers addressed by Safe Harbor

de-identification. Under this method, all dates must be generalized to year and zip codes reduced to three digits. The same approach is used on the data regardless of the context. Even if the information is to be shared with a trusted researcher who wishes to analyze the data for seasonal variations in acute respiratory cases and, thus, requires the month of hospital admission, this information cannot be provided; only the year of admission would be retained.

While important information may be lost with Safe Harbor, true de-identification can require going beyond the 18 specified identifiers. One identifier that is not mandated to be masked by Safe Harbor is occupation. In this case, someone that has a unique occupation, such as a mayor, can very easily be identified. Even though Safe Harbor does not require expert know-how and is relatively simple to implement, it is criticized as being too rigid in how data gets de-identified. Not only can this method cause valuable information to be lost, it does not ensure that a person could not be identified from their data. It is also worth noting that Safe Harbor does not meet the de-identification standards of other jurisdictions and, therefore, cannot be used outside of the U.S.

An Overview of Expert Determination

Expert Determination takes a risk-based approach to de-identification that applies current standards and best practices from the research to determine the likelihood that a person could be identified from their PHI. This method requires that a person with appropriate knowledge of and experience with generally accepted statistical and scientific principles and methods render the information not individually identifiable[4]. It requires:

1) 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 a determination.

Expert Determination is considered a risk-based approach because the amount of de-identification that gets applied to the data is based on an assessment of the risks related to its use or disclosure. A dataset containing highly sensitive information, such as the results of a drug test, would not be treated in the same way as a dataset containing the names of newspaper subscribers. The former would

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

have more rigorous de-identification applied so that it would take an exceptional level of skill, effort and resources to successfully re-identify individuals within it.

Expert Determination does not have the same downsides as Safe Harbor and, therefore, is the recommended method for de-identification by numerous highly regarded organizations including the Institute of Medicine, HITRUST and the Canadian Council of Academies. In addition, because it is based on statistical principles, it employs techniques that can be used to automate the de-identification process with software. This method can also be applied internationally, as it is in line with legislation in other parts of the world.

Data custodians need to understand both the content of their datasets and the context in which they will be shared to ensure their data is de-identified effectively. A balance must be sought between maximizing the privacy of personal information and maximizing the usefulness of the data. Achieving this is done by applying a combination of data masking and de-identification techniques to a dataset, including pseudonyms, suppression, sub-sampling and aggregation.

The HIPAA Privacy Rule lays out two standards for the de-identification of PHI: Safe Harbor and Expert Determination. While Safe Harbor is relatively simple to implement it has drawbacks that reduce the usefulness of the data and potentially leave it open to re-identification. Expert Determination offers a more robust and comprehensive approach to de-identification that is based on probable risks and requires the expertise of individuals who are knowledgeable in the scientific methods and principles of de-identification. This enables Expert Determination to serve as a strong foundation for an automated de-identification process.

Continue the journey by reading the next paper in this series *De-identification 301: Three Adversaries Who Could Attack Your Data.*

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.
Anonymization	Sometimes used interchangeably with the term 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.
Business Associate	Business Associates are defined under HIPAA as a person or entity that performs certain functions or activities that involve the use or disclosure of protected health information on behalf of, or provides services to, a Covered Entity.
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.
Covered Entity	Covered entities are defined under HIPAA as health plans, healthcare clearinghouses and healthcare providers that electronically transmit any health information. By law, the HIPAA Privacy Rule applies only to Covered Entities.
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	See Anonymization. The term de-identification is used more frequently in the United States.
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.
Generalization	See Aggregation.
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).
HIPAA Privacy Rule	Sets out the standard for privacy of individually identifiable health information Contained within the Health Insurance Portability and Accountability Act, the HIPAA Privacy Rule applies to organizations that are defined as Covered Entities under the Act and requires that those that work with HIPAA Business Associates produce a contract that imposes safeguards on the PHI that the business associate uses or discloses.
Indirect Identifier	Also referred to as a 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.

Appendix: Terminology

Term	Definition
Masking	A process that reduces the risk of identifying a data subject to a small level by applying a set of data transformation techniques without concern for the analytic value of the data.
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.
Personally Identifiable Information (PII)	Data that can be used to uniquely identify or locate an individual. Examples of personally identifiable information include name, phone number or credit card number.
Quasi-Identifier	See Indirect Identifier.
Record Suppression	Removing a record from the dataset when the combination of quasi-identifiers presents a high risk of re-identification.
Re-identification	The identification of a unique individual within a dataset that was supposed to have been de-identified.
Safe Harbor	A standard methodology for de-identification specified under the HIPAA Privacy Rule. The Safe Harbor methodology requires the removal of 18 types of direct and quasi-identifiers from a dataset so that no actual residual information can be used to identify an individual.
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.
Statistical Method	See Expert Determination.
Sub-Sampling	Taking a random sample of records from a dataset. For example, if the requirement is for a dataset of 1000 records, this could be achieved by taking a random sample of 10% of the records from a larger 10,000 record dataset.

Sources

1 Pepitone, Julianne. 5 data breaches: From embarrassing to deadly. CNN Money. Retrieved from http://money.cnn.com/galleries/2010/technology/1012/gallery.5_data_breaches/

2. Sweeney, L. (2000). Simple Demographics Often Identify People Uniquely. Data Privacy Lab Identifiability Project. Retrieved from http://dataprivacylab.org/projects/identifiability/index.html

3. Department of Health and Human Services (2002, August 14). Standards for Privacy of Individually Identifiable Health Information; Final Rule. 45 CFR Parts 160 and 164. Retrieved from http://www.hhs.gov/ocr/privacy/hipaa/administrative/privacyrule/privruletxt.txt

4. Department of Health and Human Services (2002, August 14). Standards for Privacy of Individually Identifiable Health Information; Final Rule. 45 CFR Parts 160 and 164. Retrieved from http://www.hhs.gov/ocr/privacy/hipaa/administrative/privacyrule/privruletxt.txt